PRINCIPLES OF ELECTRICITY

AND

ELECTROMAGNETISM

INTERNATIONAL SERIES IN PURE AND APPLIED PHYSICS

G. P. HARNWELL, *Consulting Editor*

BACHER AND GOUDSMIT—ATOMIC ENERGY STATES
BITTER—INTRODUCTION TO FERROMAGNETISM
BRILLOUIN—WAVE PROPAGATION IN PERIODIC STRUCTURES
CADY—PIEZOELECTRICITY
CLARK—APPLIED X-RAYS
CURTIS—ELECTRICAL MEASUREMENTS
DAVEY—CRYSTAL STRUCTURE AND ITS APPLICATIONS
EDWARDS—ANALYTIC AND VECTOR MECHANICS
HARDY AND PERRIN—THE PRINCIPLES OF OPTICS
HARNWELL—ELECTRICITY AND ELECTROMAGNETISM
HARNWELL AND LIVINGOOD—EXPERIMENTAL ATOMIC PHYSICS
HOUSTON—PRINCIPLES OF MATHEMATICAL PHYSICS
HUGHES AND DUBRIDGE—PHOTOELECTRIC PHENOMENA
HUND—HIGH-FREQUENCY MEASUREMENTS
PHENOMENA IN HIGH-FREQUENCY SYSTEMS
INGERSOLL, ZOBEL, AND INGERSOLL—HEAT CONDUCTION
KEMBLE—PRINCIPLES OF QUANTUM MECHANICS
KENNARD—KINETIC THEORY OF GASES
KOLLER—THE PHYSICS OF ELECTRON TUBES
MORSE—VIBRATION AND SOUND
PAULING AND GOUDSMIT—THE STRUCTURE OF LINE SPECTRA
RICHTMYER AND KENNARD—INTRODUCTION TO MODERN PHYSICS
RUARK AND UREY—ATOMS, MOLECULES AND QUANTA
SCHIFF—QUANTUM MECHANICS
SEITZ—THE MODERN THEORY OF SOLIDS
SLATER—INTRODUCTION TO CHEMICAL PHYSICS
MICROWAVE TRANSMISSION
SLATER AND FRANK—ELECTROMAGNETISM
INTRODUCTION TO THEORETICAL PHYSICS
MECHANICS
SMYTHE—STATIC AND DYNAMIC ELECTRICITY
STRATTON—ELECTROMAGNETIC THEORY
WHITE—INTRODUCTION TO ATOMIC SPECTRA
WILLIAMS—MAGNETIC PHENOMENA

Dr. Lee A. DuBridge was consulting editor of the series from 1939 to 1946.

PRINCIPLES OF ELECTRICITY

AND

ELECTROMAGNETISM

BY

GAYLORD P. HARNWELL

Mary Amanda Wood Professor of Physics, University of Pennsylvania

SECOND EDITION

NEW YORK TORONTO LONDON

McGRAW-HILL BOOK COMPANY, INC.

1949

PREFACE TO THE SECOND EDITION

The ten years that have intervened since the publication of the first edition of this book have seen great advances in electronics and in atomic and nuclear physics. Although these topics are not actually in the field of electricity and electromagnetism, they have much in common with it. Significant developments in them influence both the physical basis of electrical theory and the emphasis to be accorded certain practical applications. As one consequence of this the accounts of electronics and radiation in the first edition have been expanded somewhat in the direction of the higher frequencies that have become important in radar and atomic-physics research. Also brief accounts of generators of these frequencies and of resonant cavities and wave guides are now included.

Most present students of electricity and electromagnetism are more familiar with atomic physics than those of a decade ago. This unifying point of view is now commonly presented in elementary physics courses. Also our knowledge of the electric and magnetic properties of atoms in various states of aggregation has grown significantly in recent years. In consequence a somewhat greater familiarity with atomic physics is assumed, and more attention is devoted in this edition to the physics of solid conduction, the magnetic properties of atoms and nuclei, and high-energy electromagnetic ion accelerators.

There have also been numerous other changes of greater or lesser moment throughout the book. The chapter on nonohmic circuit elements has been largely rewritten both to take account of the development of thermistors and to endeavor to improve the earlier presentation. Certain errors in the previous edition have been corrected, and the author is greatly indebted to his many friends and users of the book for pointing these out to him. The assistance of Professors J. Halpern, P. H. Miller, Jr., W. E. Stephens, and C. Weygandt, through helpful criticism and constructive suggestion, is gratefully acknowledged.

GAYLORD P. HARNWELL

PHILADELPHIA, PA.
January, 1949

PREFACE TO THE FIRST EDITION

The physical phenomena that are grouped together under the heading of electricity and magnetism are of basic importance in almost every branch of science. The fields of electrical power and communication engineering are the ones of greatest economic moment, but electrical devices and electrical measuring techniques are valuable tools in all the branches of natural science and in medicine. Fundamental electrical phenomena are, however, primarily the concern of the physicist, and it is from the point of view of the experimental physicist that this account of the elements of the subject has been prepared. It is intended as an introduction to both experimental and theoretical electricity, but the emphasis is placed on the experimental aspect rather than the theoretical. In addition to the fundamental classical phenomena, elementary discussions of electronics and gas discharges have been included, as many of the most interesting modern developments are taking place in these fields.

Though the subject is here developed from the basic experimental laws, these are introduced as being to some extent familiar to the student. It is assumed that he has had a sound introduction to physics in an elementary course and that this is his second encounter with the concepts of electricity. A general familiarity with the subjects of mechanics and heat, such as would be gained from an introductory course in physics, is also presupposed. Though the theoretical development is not the primary purpose of this text, it is essential to an understanding of the subject, and it has been necessary to assume a certain knowledge of elementary mathematical techniques. This, however, is limited to a knowledge of the differential and integral calculus and elementary differential equations. One of the appendices is a résumé of the types of first- and second-order differential equations that are most frequently encountered in the text. Use is made of the vector notation, as this was developed primarily for dealing with physical phenomena such as electricity and magnetism. It simplifies the presentation and facilitates the statement of the fundamental laws in a general form independent of special coordinate systems. This concise formulation makes the expressions more easily remembered and reduces the mathematical typography to a minimum. For the benefit of students who have not previously employed this technique the necessary concepts and notations are developed *ab initio* in an appendix.

As this presentation is concerned primarily with the experimental aspect of the subject, the system of units that has the widest practical acceptance has been adopted throughout. This is the system in which the unit of work is the joule, the unit of power the watt, the unit of charge the coulomb, and the unit of resistance the ohm. This system of units is used almost exclusively in all practical applications, and, to avoid the confusion which frequently results from the use of different units in different parts of the subject, this system is adopted at the beginning and used consistently. Other possible systems are, of course, discussed and used explicitly from time to time to familiarize the student with the classical electrostatic and electromagnetic systems. The mechanical units chosen are the meter, kilogram, and second.[1] A consistent absolute system of electrical units can be based upon this choice, and the units of length and mass are the actual international standards rather than subdivisions of them. Furthermore, the International Committee of Weights and Measures has adopted the *absolute system* of units based on this choice of the fundamental quantities.

The International Committee has not chosen any particular fundamental electrical unit. In this treatment the permeability of free space, which is written μ_0, will be chosen as exactly $4\pi \times 10^{-7}$ henry per meter. The factor 4π, which classes these units as "rational," is included in order to simplify the electromagnetic relations rather than Coulomb's law of electrostatic attraction. The reason for the adoption of rationalized units is that the electromagnetic relations are more frequently used than the electrostatic equation in most applications. It is unfortunate from the pedagogic point of view that Coulomb's law in these units involves a constant and lacks the simplicity associated with the quantitative law in the electrostatic system. The necessity for a constant of proportionality in this law is inherent in the absolute practical system of units and rationalization merely increases it by the factor 4π. The situation is somewhat analogous to that of the universal law of gravitation, which contains a dimensional constant since the newtonian dynamical law, $F = ma$, has been chosen without one. The above dynamical law, of course, holds in this form in the system of units here adopted and the name given the unit of force, which will give a mass of 1 kg. an acceleration of 1 m. per second per second, is the *newton*.

The constant in Coulomb's law of electrostatics has a certain advantage in that it conserves the identity of the electrical and mechanical units on the two sides of the equation. As the force is in newtons, the charges in coulombs, and the distance in meters, the constant of propor-

[1] It should be pointed out that density and specific gravity are not the same in this system of units. The maximum density of water is evidently 1,000 Kg. per cubic meter and that of dry air at 0°C., and 76 cm. Hg is 1.2928 Kg. per cubic meter.

tionality that occurs in the denominator has the dimensions coulombs squared per newton per meter squared, and no question need arise as to the dimensions of charge in the fundamental mechanical units. Furthermore, this constant appears in the relation between electric field and electric polarization and displacement. In consequence of its introduction the field is measured in terms of volts per meter and the other two quantities in terms of coulombs per square meter. This serves to draw an additional valuable distinction between these entities. A similar situation arises in connection with the magnetic field and magnetization and induction that serves to differentiate these quantities in the mind of the student.

In this treatment electric charge is considered as the fundamental electric and electromagnetic entity. It is the additional datum that must be introduced to extend the laws of mechanics to cover electric and electromagnetic phenomena. The first division of the subject that is taken up is that of electrostatics, which deals with the configurations of charges at rest and the resultant mechanical forces between them. The physical concepts involved in this field are simpler than those of electromagnetism. This division also serves to introduce the vectorial concepts of gradient and divergence in their most evident physical form. In this classical field the examples have been limited to the simplest geometries of points, planes, cylinders, and spheres. The nonmagnetic phenomena that are associated with the flow of electric charge are then considered. It is assumed at this point in the text that the student is familiar with the basic electrical instruments such as the galvanometer and its derivatives, the voltmeter and ammeter. The principles of operation of these instruments are not taken up until a later chapter. In view of the familiarity of the average student with alternating currents at power frequencies, this topic is introduced in connection with linear and nonlinear resistances, which is somewhat earlier than in the usual treatment. Also the consideration of nonohmic circuit elements and the conduction of electricity in gases that follows is somewhat more extensive than is common in a general text. It is felt that their importance in modern circuit theory justifies this emphasis.

The theory of magnetism is approached from the point of view of the electromagnetic phenomena associated with moving charges rather than from that of permanent magnetism. While the formal presentation in this way is not so direct, the fundamental unity of the subject is kept more clearly in mind. There is an economy of hypotheses, and there is also less tendency to confuse the significance of the different magnetic vectors. The concept of the magnetic pole is introduced, and the formal analogy between electrostatics and magnetostatics is indicated as an aid in calculation. However, the magnetic pole is regarded as a useful formal analogy rather than as an essential physical entity that is required

for the adequate description of experimental phenomena. The choice of illustrative material in the field of modern alternating-current-circuit theory has been guided largely by the importance of the various circuit elements and their combinations in the field of pure scientific research and in that of communication engineering. The inclusion of the latter field is due to the widespread interest in the theory of radio communication on the part of a large group of students who have become more or less familiar with the practical details of the subject as ardent radio amateurs. The last division of the subject that is considered is that of electromagnetic radiation. In this discussion the emphasis has been laid on the frequency range that is generated with electrical circuits rather than on the higher range, which may be more properly considered as the province of physical optics.

It is with some trepidation that a book differing as widely as this does from much precedent in a well-established field is offered to students and teachers of electricity. However, the subject has developed and expanded far beyond its position when much of the classical precedent was established, and it is felt that it is no longer adequate to consider the subject primarily from a classical point of view without a fundamental integration of the modern theoretical outlook and the interesting and important technical advances. The author's obligations to the many previous treatments of the subject are obvious throughout the text and are too numerous for individual mention. It is a great pleasure to acknowledge Professor H. P. Robertson's helpful criticism of the problems of mathematical presentation and that of Professor C. W. Willis in connection with the brief treatment of the technical material in Chap. XII. In addition, the author is greatly indebted to Dr. S. N. VanVoorhis and Dr. J. B. H. Kuper and to his colleagues at Princeton for their valuable advice and suggestions.

GAYLORD P. HARNWELL

UNIVERSITY OF PENNSYLVANIA
September, 1938

CONTENTS

CHAPTER V

CHAPTER VI

CHAPTER VII

CHAPTER X

CHAPTER XI

CHAPTER XII

CHAPTER XIII

CHAPTER XIV

CHAPTER XV

CHAPTER XVI

MATHEMATICAL APPENDIX

PRINCIPLES OF ELECTRICITY AND ELECTROMAGNETISM

CHAPTER I

ELECTROSTATICS

1.1. Qualitative Electrostatic Phenomena.—All our senses yield information, either direct or indirect, regarding the group of physical phenomena that are classed together under the general heading of electricity and magnetism. Probably our most immediate experience of electricity is produced when small currents flow through the body and produce a tingling sensation. However, sensations of this type are not suitable as a basis for a quantitative study of electricity. The only experimental measurements that can be made in a strictly quantitative way are those involving the position and change in position, or displacement, of a physical object. Measurements of this type constitute the data upon which the science of mechanics is based. A standard of length is defined and other lengths compared with it by means of a series of measurements of position. Similarly time is measured by the positions of the heavenly bodies or by the positions of clock hands, and after a standard of mass has been defined, other masses are compared with it by noting the null position of a balance pointer. It is only by making measurements of this type and correlating the results with mechanical laws that the phenomena of electricity can be studied in a quantitative manner. The derived mechanical concepts of force and work play an important role in this study, and it was through a series of qualitative observations on the forces that can be brought into existence by certain procedures that the fundamental concepts of electricity were first developed.

The early Greek philosophers were aware of the fact that when amber, which is a yellowish resin, is rubbed it acquires the property of attracting to itself small bits of matter. That is, forces are brought into existence between these particles and the amber which are sufficiently strong to counteract, for instance, the force of gravity. The Greek word meaning amber is "ἤλεκτρον" and it is from this that our word electricity is derived. When amber or any other material is in such a condition that it gives rise to these forces, it is said to be *electrified.* The production of electrification by the frictional process of rubbing is known as *triboelectrification.* William Gilbert, in the sixteenth century, investigated

this type of electrification and on the basis of his observations divided all materials into two classes: Substances such as glass, amber, silk, etc., which he was able to electrify he called "electrics" and other substances such as the metals which he was unable to electrify he called "non-electrics." These two groups we now know as *insulators* and *conductors.* About a century later, Du Fay showed that the distinction between these groups depends not on their ability to become electrified but on their ability to retain electrification at the point where it is produced. The condition of electrification is lost by a conductor if it is connected through a conducting path with some other large conductor such as the earth.

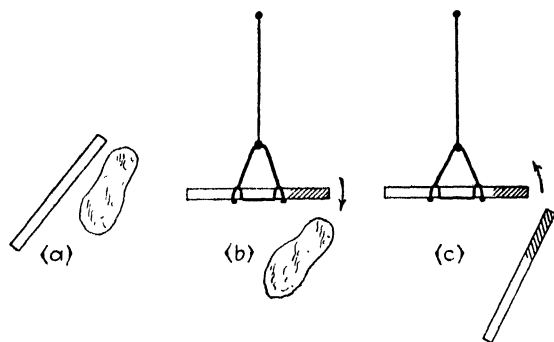



FIG. 1.1.—(*a*) Hard-rubber rod and a piece of wool with which it is to be rubbed, between which there is initially no force. (*b*) Force of attraction between the wool and the end of the hard-rubber rod against which it was rubbed. (*c*) Force of repulsion between the ends of two hard-rubber rods that have both been rubbed with wool.

Du Fay also found that it is necessary to assume two types of electrification in order to explain the phenomena he observed. These two types he called "vitreous" and "resinous" because the first was associated with glasslike and the second with resinlike materials. A few years later Benjamin Franklin suggested that there was really only one type of electricity that was present to a certain normal extent in all uncharged or neutral bodies. An excess over the normal quantity of this electricity which he called positive (+) gave rise to the phenomena of vitreous electrification and a lack of electricity produced negative (−) or resinous electrification. We find it more convenient at our present stage of understanding of electrical phenomena to retain the concept of two types of electricity as suggested by Du Fay. These types are present in equal quantities in a neutral or uncharged body. If this balance is disturbed, the body is said to be charged and we use the terms positive and negative as being particularly adapted to algebraic manipulation.

A few simple instances of the types of experiments upon which these ideas rest will be discussed briefly. If one end of a hard-rubber rod is rubbed briskly with a piece of dry wool and is then suspended in a cradle at the end of a long fiber, as shown in Fig. 1.1, it will be found that when

the wool is brought near the rubbed end of the rod, the attractive force between the two will cause the charged end of the rod to swing toward the wool. If another hard-rubber rod is rubbed in a similar manner and is then brought up toward the charged end of the suspended rod, the latter will be repelled and will swing away from the second rod. If any other material is electrified by friction and approached to the suspended rod, the latter will be either attracted or repelled. If it be arbitrarily assumed that the charge on a rubber rod after being rubbed with wool is negative, it is evident that two negative charges repel one another. The positive sign produced by the product of two negative ones by the rules of algebra is associated with a force of repulsion. Substances that attract the rod are said to have positive charges, and the negative sign resulting from the product of positive and negative is associated with a force of attraction. If a positively charged rod is suspended in the cradle, it is found to be repelled by other positive charges and attracted by negative ones in accordance with the algebraic rules proposed above. A substance does not always acquire a charge of the same sign on being rubbed with different materials. However, the following rule is found to hold. If rubbing together materials A and B makes A + and B − and if rubbing materials B and C makes B + and C −, then if materials A and C are rubbed together, A will become + and C −. Hence substances that can be tested in this way can be classified in a table such that a substance will be electrified positively on being rubbed with another if the second substance occurs below it in the table, and will be electrified negatively if the second substance occurs above it. Such a table is known as a *triboelectric series.* The relationships which it expresses are at best approximate as the position of a substance is determined to some extent by other factors such as temperature, humidity, etc.

TRIBOELECTRIC SERIES
(*Smithsonian Tables*)

Rabbit's fur
Glass
Mica
Wool
Cat's fur
Silk
Cotton
Wood
Amber
Resins
Metals (Cu, Ni, Co, Ag, etc.)
Sulphur
Metals (Pt, Au)
Celluloid

As a result of modern experiments in atomic physics we have reason to believe that the atoms which constitute the different chemical elements and which in a state of aggregation form the molecules characteristic of chemical compounds are themselves composed of positive and negative charges. The mass of the atom is largely associated with the positive charges and certain neutral constituents; these together form a central nucleus, the net positive charge of which determines the atomic species or chemical element. Surrounding this nucleus is a cloud of negative charges which possesses a characteristic configuration. For a normal atom the magnitude of this negative charge is equal to that of the net positive charge on the nucleus with the result that the atom is electrically neutral. This extranuclear cloud of charge is composed of an integral number of elementary indivisible negative charges known as *electrons*. The mass of an electron is less than that of a nucleus by a factor of the order of 10^4. The electron configuration is maintained by forces which are at least partially of the nature of those between macroscopic charges. When an electron is removed from a normal atom, the result is a positively charged particle known as a *positive ion*. The work necessary to remove an electron from an atom can be measured and it is found that certain electrons of the configuration are much more loosely bound to the positive nucleus than others. In the solid state of aggregation the positive nuclei and their surrounding electron configurations are held in fixed relative positions by certain forces and it is to this rigid spacial structure that the permanence of shape of a solid body is ascribed. In certain types of solids the space lattice formed by the positions of the constituent atoms is such that one or two of the more loosely bound electrons associated with each atom are not held rigidly to it but are so influenced by the neighboring electric charges of the solid that they are able to move more or less freely through the lattice. They are, however, retained by the physical boundaries of the solid as by the walls of a box. These mobile electrons are known as *conduction electrons*. The types of solids that exhibit phenomena which can be understood at least to a first approximation by a qualitative picture of this sort are known as conductors. If all the constituent electrons of a solid are bound rigidly to the atoms forming the solid lattice structure so that no relative motion of the charges is possible, the phenomena to be expected would be those characteristic of the class of substances known as insulators. For with these, if a positive or negative charge is produced by friction in a particular region, it remains in that locality and does not distribute itself over the object as it would in the case of a conductor.

This qualitative picture will be amplified and made more definite and quantitative in later discussions, but it is of assistance at this point in understanding certain of the simple phenomena of electrification in

insulators and conductors. Consider, for instance, a glass rod that is charged positively by rubbing with silk. In some way that is little understood electrons that were previously associated with the surface atoms of the glass are effectively rubbed off and adhere to the silk. A light sphere of fiber or pith wrapped with thin aluminum foil and suspended by an insulating silk thread is frequently used as a test body. If the rod is brought in contact with the pith ball, electrons will flow from the foil to neutralize the charge at the point of contact, and if the rod is then removed, the ball is left positively charged at the expense of part of the charge previously on the rod. The rod will then repel the pith ball if brought toward it. If a conducting body on an insulating support is placed with one end near the pith ball and the rod approached to its other end, as illustrated in Fig. 1.2, the pith ball will be repelled as from the proximity of the rod itself. The conduction electrons, being negative, are attracted toward the end of the conductor near the rod, leaving a net positive charge on the far end, under the influence of which the positively charged pith ball is repelled. Thus the influence of the neighboring charges on the rod effectively separates a fraction of the charges on the neutral conductor. This phenomenon is known as *induction*.

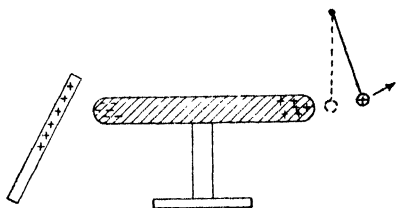



FIG. 1.2.—Transmission of the mechanical effect of an electric charge, or the separation of the charges on a neutral conductor under the influence of a neighboring charge (induction).

It is possible to demonstrate the presence of charge quite directly by dusting over the body a mixture of powdered red lead and sulphur that has been previously shaken together. During the shaking the sulphur becomes negatively charged, and the red lead is chargedpositively, by friction between the grains; when the mixture is dusted over a body the sulphur adheres to positively charged areas and the red lead to those that are negatively charged, imparting characteristic colors to the charged regions.

The electroscope, which is an instrument for determining the magnitude and sign of an electric charge, is a further illustration of induction. One form of the instrument is shown diagrammatically in Fig. 1.3. A metal container with an open top is supported from the bottom by a metal rod that runs down through a cork into a glass flask. Attached to this rod a few inches above its lower end is a light conducting fiber or strip of thin gold or aluminum leaf. Normally, when no charges are present, the leaf or fiber hangs down in contact with the rod. But if, for instance, a positive charge is brought near the conducting system, as indicated at the left in Fig. 1.3, the conduction electrons tend to move up toward it, leaving a net positive charge on the leaf and rod in its neighborhood. As these positive charges repel one another, the

leaf is pushed out against the pull of gravity and the displacement of the leaf is a measure of the magnitude of the charge. If the positively charged rod is now touched to the conducting system, some of the electrons leave the latter and the resultant net positive charge causes the leaf to continue to diverge even after the charged rod is removed. If now a negative charge is brought up, conduction electrons are repelled toward the leaf, neutralizing the positive charge and causing the leaf to fall. A positive charge would produce the reverse effect and thus the sign of a charge can be distinguished. The human body is a sufficiently good conductor so that if a neutral electroscope is near a positively charged rod and the conducting system of the electroscope is touched with a finger, electrons, attracted by the charge, will flow from the body to the instrument. If the finger is then removed before removing the positive charge, the instrument will be left with a net negative charge. This is known as charging by induction; it results in the acquisition of a charge of the opposite sign to that originally brought up to the electroscope.

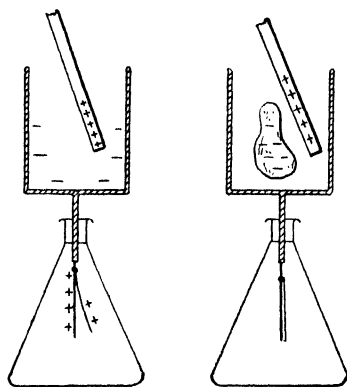



FIG. 1.3.—Electroscope demonstration that equal amounts of positive and negative charge are separated at a charging process.

In accordance with the qualitative theory of these effects that has been presented it is evident that electrification consists in the separation of charge rather than in its production. The fact that this is so and that equal amounts of positive and negative electricity always appear at any charging process can also be demonstrated with the electroscope. If the instrument is originally uncharged and a glass rod that has been rubbed with a piece of silk is placed in the upper container, the leaf will be seen to diverge. But if the silk is also placed in the container, as indicated at the right in Fig. 1.3, the leaf will fall to its normal position, indicating that there is no net charge in the container. This is a demonstration in a simple instance of what is a very important and fundamental general electrical law. No experiments have ever been performed that indicate that electrical charge can be either created or destroyed. Thus we may state that in any process electrical charge is always conserved. This law ranks with that of the conservation of momentum and with that of the conservation of mass or energy.

Machines can be built for the separation of charge that are continuously operating. These are known as static machines or electrostatic generators. These machines operate on the principle of induction as illustrated by a simple piece of demonstration equipment known as

the *electrophorus*. This consists of a flat wax surface and a conducting plate with an insulating handle. The wax is first rubbed with some material that leaves, say, a positive charge on its surface. If the plate is then set upon the wax, a small fraction of the conduction electrons from the plate will flow to the points of contact with the wax, but there are actually very few points of contact and the major effect is the separation of charge on the plate as indicated in Fig. 1.4. If a contact is then made between the plate and ground, electrons will be drawn to the plate by the positive charge on the wax, and when the contact is removed, the plate is left with a net negative charge. The insulating handle may then be grasped and the plate removed, carrying with it the negative charge to any desired place. If this charge is removed by contact with some other object the plate may be returned to the wax and the process

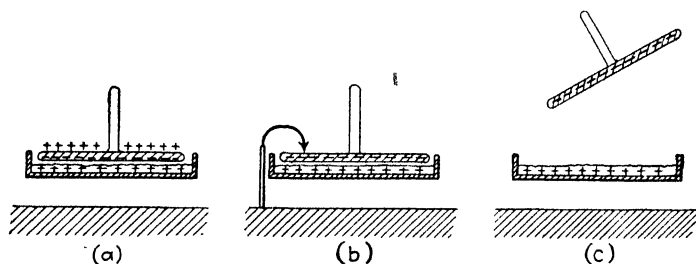



FIG. 1.4.—The electrophorous. (*a*) Neutral metal disk laid on a positively charged wax surface. (*b*) Excess positive charge is conveyed to the earth through a conducting connection. (*c*) A net negative charge is retained by the disk.

repeated. In this way an indefinite amount of charge may be transferred to any object by repeated contacts with the plate.

In the Wimshurst type of generator rotating glass disks carry what are effectively electrophorus plates and produce this transfer of charge in a continuous process. The machine is illustrated diagrammatically in Fig. 1.5. One disk is represented as being larger than the other for the purposes of the diagram and the shaded portions represent patches of conducting metal foil attached to the disks. The arms S_1 and S_2 are of metal, insulated from one another, and terminated in light wire brushes that rub over the metal patches as they rotate. The other contacts to the disks which lead up to the spherical terminals marked + and − are of the form of sharply pointed metal combs between which the conducting patches pass. The action of these is somewhat the same as that of the brush contacts. If the surface beneath the comb is, say, positively charged, electrons are drawn to the points of the comb. Under the influence of the strong electrical forces near the tips of the comb the surrounding air becomes effectively conducting. The positive and negative charges constituting the air molecules are separated; the

positive charges pass to the comb, neutralizing its charge, and the negative ones pass to the conducting patch. This results effectively in a transfer of electrons from the points of the comb to the conductor beneath it. This process is known as *corona* transfer and is widely used in electrostatic devices in air.

Assume that at the start there is a small residual positive charge on a_1. This will induce a negative charge on b_3 and a positive one on β_3. When b_3 moves into the position b_1, it induces a positive charge on the conducting patch which is then in the position a_3 and a negative charge on the conductor then in the position α_3. This process builds up positive charges on the plates in the a and β positions and negative charges on those in the b and α positions. When these charges achieve a sufficient magnitude, they are partially drawn off by corona to the combs. The a's and β's passing under the right-hand comb deliver a positive charge to it and the b's and α's deliver a negative charge to the left-hand comb. Thus the spherical terminals will become highly charged in the sense indicated in the diagram. A limit to the magnitude of the charge is set by corona from these electrodes or the occurrence of a disruptive spark between them.

FIG. 1.5.—Wimshurst type of electrostatic generator.

The Van de Graaff type of generator is superior to the Wimshurst machine for the separation of large quantities of charge. A large hollow metal sphere is supported on an insulating column, as indicated diagrammatically in Fig. 1.6. A belt of some insulating material such as silk or rubber runs in the column between a motor-driven pulley at the base and an idling pulley up inside the sphere. This acts as a conveyor belt carrying, say, positive charge to the sphere and negative charge from it to the ground. Assume that the lower pointed electrode between the belts has a residual negative charge or that the outer electrode is charged positively by some external source such as the battery indicated in the diagram. The ascending belt is then charged positively by corona from the electrode e_1 and on reaching e_3 is discharged by corona making this electrode positive. e_4 then becomes negative by induction, spraying the downward moving portion of the belt with negative charge which is then partially transferred to the electrode e_2 which tends to enhance the positive charge gained by the ascending belt. Thus the rate of transfer of charge increases as the process proceeds. Negative charge is lost to the ground through the pulley P_l and the electrode e_1 while

positive charge is gained by the sphere through the upper pulley P_u and the electrode e_4. The mutual repulsion of the positive charges distributes them on the outside of the sphere, leaving the electrode e_4 in a condition that is always receptive to more charge. The large radius of curvature and smooth surface of the sphere minimizes loss by corona and permits the accumulation of a very large charge upon it.

1.2. The Law of Force between Elementary Electrostatic Charges.—In order to handle electrical phenomena quantitatively it is necessary to know how the force between charges varies with the physical circumstances. It must be specified first of all that electrostatics deals with charges that are at rest with respect to the observer. If any relative motion takes place, other forces are brought into play and these will be discussed later under the heading of electromagnetism. Though the effects of forces between elementary electrical particles (electrons and ions) can be observed, the simpler experiments are performed by observing the effects of electrical forces on relatively large charged bodies. In order that the electrical charge shall be effectively localized at points in space the linear dimensions of these test bodies must be small in comparison with their separation from one another. The forces between such bodies are found to depend to some extent on the surrounding medium, but the difference in the forces between two charged bodies in air and the same two bodies in a vacuum is only of the order of 6 parts in 10,000. This difference is so small that for many purposes it can be neglected and the discussion of the effect of an intervening medium will be postponed until the following chapter. The only other factors that are found experimentally to influence the force between two charges localized at points are the magnitudes of the charges and their distance apart.

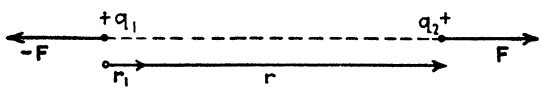
FIG. 1.6.—Van de Graaff type of electrostatic generator.

It was shown by Coulomb in 1785 that the force between two small charged bodies varies directly as the magnitude of each of the charges and inversely as the square of their separation. Figure 1.7 illustrates the principle of the method he employed. A small body carrying a charge, say q_1, is at the extremity of one arm of a torsion balance. Another small body carrying the charge q_2 is fixed in position on the circle that would be described by the rotation of q_1 about the torsion fiber. The force between these charges causes the torsion arm to rotate.

The magnitude of this force may be measured by the angle through which the knob K that supports the upper end of the fiber must be rotated in order to return the arm to its equilibrium position in the absence of the charge q_2. Thus the torque produced by the electric force between the charges is balanced by the torque of the fiber which is induced by the relative rotation of the two ends. The relation between the relative rotation of the ends of the fiber and the torque thus induced is determined by an auxiliary experiment. If by moving q_2 the distance between the charges is increased by a factor m, it is found that the torque must be decreased by the factor $1/m^2$ in order not to change the position of P_2. Next, without altering q_1 let a series of charges of different magnitude occupy the position q_2, and let the torques necessary to keep the pointer P_2 fixed be recorded. If it is now arranged that pairs of these charges replace the initial q_1 and q_2, it is found that the torques necessary to keep the pointer P_2 at the same position are proportional to the products of the torques observed when the charges separately influenced the initial q_1. Thus the force between the charges which is proportional to the torque is proportional to the magnitude of each charge and inversely proportional to the square of their separation. Writing this in the form of an equation,

FIG. 1.7.—Schematic representation of an apparatus for testing Coulomb's law.

$$F = \text{const.} \times q_1 \times q_2 \times \frac{1}{d^2} \tag{1.1}$$

This is known as *Coulomb's law of electrostatic force*. Experiments of this type are not capable of a very high accuracy. But this law of force was tested indirectly by Kelvin and Maxwell and recently by Plimpton and Lawton[1] who have shown that the exponent of d is accurately 2 to within 1 part in a billion (10^9).

Equation (1.1) is the basic quantitative law of electrostatics and the subsequent discussion will be based entirely upon it. Since F and d are mechanical quantities for which units have been previously chosen this equation may be used to define a unit of charge. The simplest procedure is arbitrarily to set the constant equal to unity and to define a unit charge by saying that if such a charge is placed a unit distance from an exactly equal charge, a unit repulsive force will be experienced by each. The *electrostatic system of units* (esu.) is defined in this way by choosing the unit of length as the centimeter and the unit of force as the dyne. Force and displacement are of course vector quantities (see Appendix *D*) and Coulomb's law may be written in vector form in these

[1] PLIMPTON and LAWTON, *Phys. Rev.*, **50**, 1066 (1936).

units to indicate both the magnitude and direction of the force. Using **boldface type** to represent vector quantities and writing $\mathbf{r}_1$ for a vector of unit length in the direction from q_1 to q_2, Coulomb's law becomes

$$\mathbf{F} = \frac{q_1 q_2}{r^2}\mathbf{r}_1$$

Here **F** is the force in dynes exerted by the charge q_1 on the charge q_2, and r is the separation of the charges in centimeters. Thus the force on a charge is in the direction of the vector to that charge if the two charges are of like sign, *i.e.*, the force is repulsive. If the charges are of unlike sign, their product has a negative sign, so **F** is in the opposite direction to $\mathbf{r}_1$ and the force is attractive. An alternative form of the

+q₁ q₂+ -F F r₁ r

$$\mathbf{F} \sim \frac{q_1 q_2}{r^2}\mathbf{r}_1$$

$$\mathbf{F} = \frac{q_1 q_2}{r^3}\mathbf{r} = \frac{q_1 q_2}{r^2}\mathbf{r}_1$$

F in dynes
r in centimeters
q in esu.
Electrostatic units

$$\mathbf{F} = \frac{1}{4\pi\kappa_0}\frac{q_1 q_2}{r^3}\mathbf{r} = \frac{1}{4\pi\kappa_0}\frac{q_1 q_2}{r^2}\mathbf{r}_1$$

F in newtons
r in meters
q in coulombs
Absolute practical units

FIG. 1.8.

law is obtained by writing **r** for the vector representing the magnitude and direction of the displacement from q_1 to q_2. In this case

$$\mathbf{F} = \frac{q_1 q_2}{r^3}\mathbf{r} \tag{1.2}$$

The electrostatic unit of charge which is defined by Eq. (1.1) with the constant set equal to unity is probably most convenient for electrostatic calculations. However, an accurate measurement of the unit charge cannot be made by a direct application of this equation and in fact electrostatic technique lacks the precision necessary for the satisfactory realization of a unit of charge. In the most widely used system of units, which is the one that will be adopted throughout this book, the unit of charge is derived indirectly through electromagnetic measurements. The rate of motion of charge is measured directly in terms of mechanical forces and the magnitude of the charge is deduced therefrom. The equations and procedures which are actually used for defining and realizing the definitions of electrical quantities will be described in Sec. 9.5. At this point it is only necessary to state that the absolute practical unit of charge is the *coulomb* which is equal to 2.998×10^9 esu. of charge. In this system of units the unit of length is the meter, the

unit of mass is the kilogram, and the unit of time is the second. Thus by Newton's law of motion the unit force is one that will impart to a mass of 1 kilogram an acceleration of 1 meter per second per second. This unit force is called the *newton.* The unit of work is the newton meter or *joule*, and the unit of power is the joule per second or *watt.* As the unit of charge is not defined through Coulomb's law on this system, a constant of proportionality will appear in the equation. As a convenience in certain types of calculations a 4π is written explicitly as part of this constant and Coulomb's law in these units becomes

$$\mathbf{F} = \frac{1}{4\pi\kappa_0}\frac{q_1q_2}{r^3}\mathbf{r} \qquad \text{or} \qquad \mathbf{F} = \frac{1}{4\pi\kappa_0}\frac{q_1q_2}{r^2}\mathbf{r}_1 \tag{1.3}$$

This is the form in which the fundamental equation will be used throughout this discussion. If the charges are measured in coulombs and if $\mathbf{r}$ is in meters and $\mathbf{F}$ in newtons, the constant κ_0 has the value

$$\kappa_0 = 8.85 \times 10^{-12}\frac{(\text{coulombs})^2}{\text{joule meter}} = 8.85 \times 10^{-12}\frac{\text{farads}}{\text{meter}}$$

to an accuracy of a few parts in 10^4. This constant, which is known as the *permittivity of free space,* is seen to have the physical dimensions of coulombs squared per joule meter, or farads per meter, where the farad is the unit of $(\text{coulombs})^2$ per joule. It is evident from Eq. (1.3) that two coulomb charges a meter apart would repel one another with a force of 9×10^9 newtons, which is of the order of magnitude of the gravitational forces of attraction between the planets.

The situation in regard to electrical units is somewhat analogous to that in mechanics in that more than one possible defining equation exists. Thus the universal law of gravitation could have been used to define a unit of force. It would then be written in the following way:

$$F = \frac{m_1m_2}{d^2}$$

That is, two unit masses a unit distance apart would attract one another with a unit force. However, it is Newton's inertial law that is actually used in mechanics to define a unit of force. This law is written

$$F = \frac{d(mv)}{dt} \qquad \text{or, if } m \text{ is constant,} \qquad F = ma$$

and a unit force is one which gives to a unit mass an acceleration of a unit distance per unit time squared. If force is defined in this way rather than by the use of the universal law of gravitation, a dimensional constant must appear in the latter law. Thus it is written

$$F = \frac{Gm_1m_2}{d^2}$$

where $G = 6.67 \times 10^{-8}$ dyne cm.^{2}/gm.^{2} in cgs. units or 6.67×10^{-11} newton meter^2 per kilogram^2, in the fundamental system of units here adopted. In the same way, if

Coulomb's law is used to define the unit of charge, any other fundamental law relating electrical and mechanical quantities will have to contain an experimentally determined constant.

1.3. Electric Field Strength and Potential.—In discussing the forces between electric charges it is convenient to introduce two additional concepts. The first of these is *electric field strength,* or more briefly electric field or electric intensity. An electric field is said to be present in a region if a small test charge placed there experiences a mechanical force. The charge that is introduced to test for the presence of a field must be infinitesimally small, ideally approaching zero in magnitude, so that it will not appreciably disturb the previously existing conditions. Here such a small charge will be called a *test charge.* The electric field at a point is defined to be the force on such a test charge at the point divided by the magnitude of the test charge, *i.e.*, it is the force per unit charge. Writing **E** for the vector representing the electric field and δq for the test charge

$$\mathbf{F} = \delta q \mathbf{E} \tag{1.4}$$

Comparing this with Eq. (1.3), it is seen that the electric field at a distance r from a point charge q is given by

$$\mathbf{E} = \frac{1}{4\pi\kappa_0} \frac{q}{r^2} \mathbf{r}_1 \tag{1.5}$$

To obtain the field due to a number of point charges it is necessary to add vectorially the fields due to the individual charges. This process can be greatly simplified by introducing the concept of *potential.* The difference in potential between two points is the mechanical work that must be expended to take a test charge from one point to the other divided by the magnitude of the test charge. The potential of a point may be thought of as the potential energy of a unit charge at the point, provided it is realized that the precise definition is the limit of the ratio of the potential energy of a test charge to the magnitude of the test charge as the latter approaches zero. If a finite charge is used to explore an electric field in the neighborhood of a conductor, the induced charges produced by the exploring charge and proportional to it will introduce effects that are not due to the previously existing field. As in the case of mechanical potential energy electric potential is determined only to within an additive constant. This constant is arbitrary and depends on the particular choice of a zero of potential. The unit of potential is the *volt.* The difference in potential between two points is 1 volt if 1 joule of work is required to transport 1 coulomb of electric charge from one point to the other. The unit of field strength is the *volt per meter.*

Consider the small vector displacement **dl** of a unit charge at a dis-

tance $\mathbf{r}$ from a single fixed charge q. The work that is done in this displacement is the product of the component of the force that opposes the motion times the distance traversed. Writing θ for the angle between the positive directions of the vector $\mathbf{E}$ and the vector $\mathbf{dl}$ as shown in Fig. 1.9, the work done against the electrostatic force is $-\mathbf{E} \cos \theta \, \mathbf{dl}$. The *scalar product* of two vectors is defined as the product of their magnitudes times the cosine of the angle included between their positive directions (Appendix *D*). It is a scalar quantity and is indicated by a dot between the vectors. Thus the work done in this infinitesimal

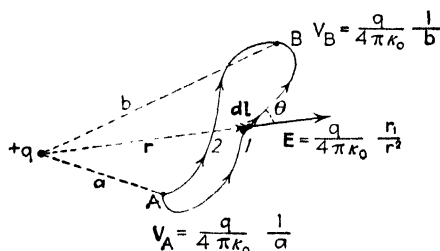



FIG. 1.9.—Work done in displacing a unit charge in the presence of a charge q.

displacement of a unit charge, which represents the infinitesimal increase in potential, is

$$dV = -\mathbf{E} \cdot \mathbf{dl} \tag{1.6}$$

But the scalar product of the unit vector $\mathbf{r}_1$ and $\mathbf{dl}$ is simply dr, the change in magnitude or $\mathbf{r}$, or from Eq. (1.5)

$$dV = \frac{-1}{4\pi\kappa_0}\frac{q}{r^2}\,dr$$

Therefore the difference in potential between the point B and the point A of Fig. 1.9 is

$$V_B - V_A = \frac{-1}{4\pi\kappa_0}q\int_A^B \frac{dr}{r^2} = \frac{q}{4\pi\kappa_0}\left(\frac{1}{r}\right)_A^B = \frac{q}{4\pi\kappa_0}\left(\frac{1}{b} - \frac{1}{a}\right) \tag{1.7}$$

From this it is evident that the difference in potential between two points depends only on the positions of the points and does not depend on the particular path that is traversed in passing from one to the other. That is, the work that is done in taking a charge from A to B via path 1 is the same that would be done if path 2 had been followed. In consequence if the charge is taken over path 1 from A to B and then returned to A via path 2, no net work is done. Writing $\oint$ for the integral around a closed path, this result may be stated

$$\oint \mathbf{E} \cdot \mathbf{dl} = (V_B - V_A) + (V_A - V_B) = 0 \tag{1.8}$$

This is the fundamental circuital potential theorem. Since the argument holds for any fixed charge q and since potential is additive, the theorem holds for any static distribution of charge whatever giving rise to the field.

If the charges that give rise to the field are located in a finite region the additive constant associated with the potential can be determined by choosing the potential zero at a very great (infinite) distance from all charges. Then the potential of a point is the work per unit charge that must be done to bring a test charge from infinity to the point in question. If A of Eq. (1.7) is the point at a very great distance, $1/a$ approaches zero and the potential V at a distance r from a point charge q is given by

$$V = \frac{1}{4\pi\kappa_0}\frac{q}{r} \tag{1.9}$$

FIG. 1.10.—The work done in taking a charge around any completely closed path in an electrostatic field is zero.

The potential due to a number of discrete charges is the algebraic sum of the potentials due to the individual charges. In the case of a continuous distribution of charge density over a volume or surface the potential is equal to the integral of Eq. (1.9) in which q is replaced by the product of the charge density times the element of volume or area. Writing q_v and q_s for volume and surface-charge densities, respectively, the general expression for V is

$$V = \frac{1}{4\pi\kappa_0}\left(\int_v \frac{q_v}{r}\,dv + \int_s \frac{q_s}{r}\,ds\right) \tag{1.10}$$

where the integration is performed over the volumes and surfaces concerned. Thus the procedure for finding the potential of a point is a simple analytical one.

The electric field at a point which determines the force that would be experienced by a charge at that point can be obtained from the potential. It has been seen that if the charges are fixed in magnitude and position, V depends only on the coordinates of the point in question. By Taylor's theorem in Cartesian coordinates a small change in potential produced by a small displacement with the components dx, dy, and dz along the three axes can be written to the first-order approximation as

$$dV = \frac{\partial V}{\partial x}\,dx + \frac{\partial V}{\partial y}\,dy + \frac{\partial V}{\partial z}\,dz$$

This may be considered to be the scalar produce of the two vectors. The first of these, which is written **grad** V and read "gradient of V," is

given by

$$\mathbf{grad}\ V = \mathbf{i}\frac{\partial V}{\partial x} + \mathbf{j}\frac{\partial V}{\partial y} + \mathbf{k}\frac{\partial V}{\partial z} \tag{1.11}$$

The second is the arbitrary vector displacement **dl** that can be written in terms of its Cartesian components as

$$\mathbf{dl} = \mathbf{i}\,dx + \mathbf{j}\,dy + \mathbf{k}\,dz$$

In accordance with the rules of the scalar product the cross products of the three unit vectors **i**, **j**, and **k** in the directions of the three axes vanish and $\mathbf{ii} = \mathbf{jj} = \mathbf{kk} = 1$, hence

$$dV = \mathbf{grad}\ V \cdot \mathbf{dl} \tag{1.12}$$

On comparing this with Eq. (1.6) it is seen that

$$\mathbf{E} = -\mathbf{grad}\ V \tag{1.13}$$

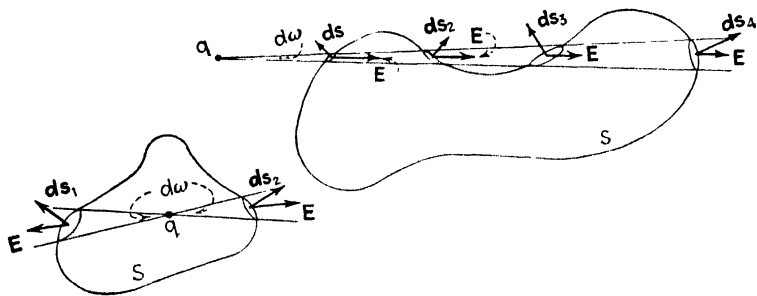
Or in terms of the components of the electric field in the directions of the three Cartesian axes

$$E_x = -\frac{\partial V}{\partial x}, \qquad E_y = -\frac{\partial V}{\partial y}, \qquad E_z = -\frac{\partial V}{\partial z}$$

Thus, if the potential of a point has been determined as a function of its coordinates, the components of the electric field there are given by the negative partial derivatives of the function at the point in question. The vector electric field is $-\mathbf{grad}\ V$ or the vector which has these components.

1.4. Gauss's Theorem for Electrostatics.—The inverse-square law of force may be stated in an alternative way that brings out certain general properties associated with the law and also facilitates many particular calculations. Consider any hypothetical closed surface and form the sum of each infinitesimal element composing the surface times the normal component of the electric field at the element. The outward normal component is reckoned as positive. If ds represents an element of the surface, the contribution to the sum made by this element is $E\,ds \cos\theta$, where θ is the angle included between **E** and the outward normal to the surface at ds. Or if **ds** is the vector representation of the surface element, *i.e.*, a vector equal in magnitude to the area ds and in the direction of the outward normal to the surface, the contribution to the sum made by this element is $\mathbf{E} \cdot \mathbf{ds}$. Writing $\oint_s$ for the integral over the whole closed surface and using Eq. (1.5)

$$\oint_s \mathbf{E} \cdot \mathbf{ds} = \frac{q}{4\pi\kappa_0}\oint_s \frac{\mathbf{r}_1 \cdot \mathbf{ds}}{r^2}$$

But each element of the integral on the right is equal to the solid angle subtended by the element ds at the point occupied by the charge q. For $\mathbf{r}_1 \cdot \mathbf{ds}$ is the product of ds and the cosine of the angle between the normal to the surface and the radius vector from q, *i.e.*, it is the projection of ds normal to the radius vector. The quotient of this normal area by r^2 is by definition the solid angle $d\omega$ subtended by ds at q. The integral over the surface is then the total solid angle subtended by the surface at the point occupied by q. It can be seen from Fig. 1.11 that if q lies outside the surface, each infinitesimal conical solid angle $d\omega$ cuts the surface an even number of times, each intersection making equal contributions of opposite sign to the sum. Thus the net contribution to the integral in this case is zero. If, however, q lies within the volume bounded by the surface, the integral of $d\omega$ over the surface yields the complete solid angle, which is the area of a sphere surrounding

$\oint \mathbf{E} \cdot \mathbf{ds} = q/\kappa_0$
FIG. 1.11.—Demonstration of Gauss's theorem.

the point divided by the square of its radius, or 4π. The argument can be applied to all the charges that are present and the result written

$$\oint_s \mathbf{E} \cdot \mathbf{ds} = \frac{q}{\kappa_0} \tag{1.14}$$

where q is the total charge contained in the volume bounded by the closed surface s. This is *Gauss's theorem* and may be stated in words by saying that the surface integral of electric field strength over any closed surface is equal to the algebraic sum of all the charges enclosed by the surface divided by the constant κ_0.

Applications of Gauss's Theorem.—Gauss's theorem may be used to obtain certain general results that are of great importance. Consider the general case of a charged or uncharged conducting body in an electric field. By definition a conductor is a substance containing mobile electrons that are free to move about throughout the volume of the conductor under the influence of an electric field. Since we are here dealing with the static case, there is no net motion of these electrons and

hence there can be no net electric field at any point inside the conductor. Therefore a test charge can be moved freely about inside a conductor without any work being done and a conductor represents a region that is all at the same potential. Thus a conductor is an equipotential volume or, if it is in the form of a thin shell, it is an equipotential surface. Consider a hypothetical closed Gaussian surface lying completely inside a conductor, as at the left in Fig. 1.12. Since the electric field at every point on this surface is zero, it can enclose no net charge. As this hypothetical surface can be expanded till it lies just beneath the surface of the conductor, there can be no net charge within the conductor itself and any charge carried by it must reside entirely on the surface.

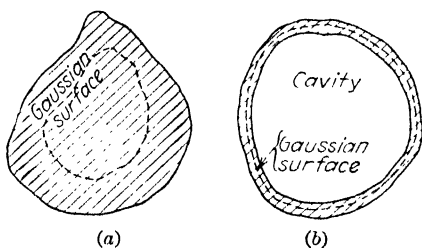



FIG. 1.12.—(a) Section through a solid conductor and a hypothetical Gaussian surface in its interior. (b) Section through a hollow conductor and Gaussian surface surrounding the cavity.

If the conductor is hollow and the Gaussian surface encloses the cavity, the theorem continues to apply. If there are no net charges inside the cavity, there can be none on the inner walls. Any net charge carried by the conductor must reside entirely on its outer surface. This is a very important result, for all the accurate tests of the inverse-square law depend on the verification of this fact. If there are no charges within the cavity, the Gaussian surface can be distorted in any manner throughout the region and $\oint \mathbf{E} \cdot \mathbf{ds}$ continues to be zero. This can be true only if the electric field vanishes at every point over the inner walls and throughout the interior of the cavity. Hence no configuration of charges external to the conductor can give rise to a field inside. The closed conducting surface completely shields the interior from any external electrostatic forces. If there is a charge, say q, inside the cavity, an equal and opposite charge $-q$ must be induced on the inner walls in order that $\oint \mathbf{E} \cdot \mathbf{ds}$ may vanish over a Gaussian surface in the conductor enclosing the cavity. In consequence if the conductor has no net charge, a charge q must appear over the outer surface. This gives rise to an external electric field and thus it is seen that the conductor does not shield the external region from the effects of an internal charge. However, it will be shown later that the external effects are independent of the position of the charge within the cavity.

Gauss's theorem may be used to obtain an equation relating the field at the surface of a conductor and the surface-charge density at the point. The field at the conductor must be normal to the conducting surface. For, if this field had a component parallel to the surface, the conduction electrons would move along the surface under the influence of

this component. Thus the distribution of charge would not be static in contradiction to the fundamental assumption. Consider a hypothetical Gaussian surface in the form of a small pillbox enclosing a portion of the surface, as shown in Fig. 1.13. The sides of the box are perpendicular to the surface, one end lies in the conductor and the other is immediately outside it and parallel to its surface. The portion of the surface inside the conductor makes no contribution to the integral of Eq. (1.14) since it has been seen that the field is zero inside the conductor. Likewise the sides of the box perpendicular to the surface make no contribution since there is no component of the field perpendicular to them. The only contribution to the integral is made by the outer cap and if this is of area ds, the integral has the value $E_n\, ds$, where E_n is the normal component of the electric field just outside the surface. If q_s is the surface-charge density, the total charge enclosed by the pillbox is $q_s\, ds$ and Eq. (1.14) becomes

$$E_n = \frac{q_s}{\kappa_0} \qquad (1.15)$$

$\mathbf{E} = q_s/\kappa_0\, \mathbf{s}_1$
$\mathbf{s}_1$ = Unit vector perpendicular to the surface of the conductor
q_s = Charge per unit area of conductor
FIG. 1.13.

This is a very important and useful relation. As an example it is found by experiment that the limiting field in air before corona takes place is of the order of 3×10^6 volts per meter. Inserting the numerical value of κ_0, this is seen to correspond to a surface-charge density of about 2.65×10^{-5} coloumb per square meter.

Gauss's theorem is particularly useful in calculating the electric field when the charge distribution presents a high degree of symmetry. Consider first an isolated sphere with a charge q uniformly distributed over its surface (a condition that would automatically be achieved if the sphere were conducting). By symmetry the external field must be radial at every point and can depend only on the distance from the sphere. If a Gaussian surface in the form of a concentric enclosing sphere of radius r is drawn and Eq. (1.14) applied $\mathbf{E}$ may be removed from the integral since it is constant and perpendicular to the surface. Hence

$$\oint \mathbf{E} \cdot \mathbf{ds} = E_r \oint ds = 4\pi r^2 E_r = \frac{q}{\kappa_0}$$

or

$$E_r = \frac{q}{4\pi r^2 \kappa_0} \qquad \text{(sphere)} \qquad (1.16)$$

On comparing this expression with Eq. (1.5) it is evident that the external field due to the charged sphere is the same as the field which would be

produced if the entire charge were concentrated at the center. The case of a uniformly charged circular cylinder of infinite length can be treated in an analogous way. Here the electric field is by symmetry perpendicular to the axis of the cylinder and can depend only on the distance r from the axis. A section of an enclosing coaxial cylinder is chosen as the Gaussian surface as shown at the right in Fig. 1.14. Since $\mathbf{E}$ is parallel to the ends of this section, they make no contribution to the integral. $\mathbf{E}$ is uniform over the curved portion of the surface

$$\mathbf{E} = \frac{q}{4\pi\kappa_0}\frac{\mathbf{r}}{r^3} = \frac{q}{4\pi\kappa_0}\frac{1}{r^2}\mathbf{r}_1 \qquad \mathbf{E} = \frac{q_l}{2\pi\kappa_0}\frac{\mathbf{r}}{r^2} = \frac{q_l}{2\pi\kappa_0}\frac{1}{r}\mathbf{r}_1$$

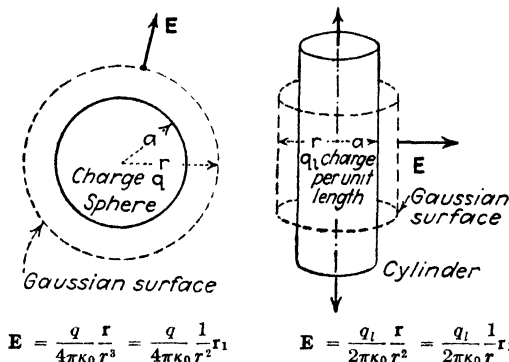



FIG. 1.14.—Gauss's theorem applied to a sphere and to a section of an infinitely long cylinder.

and hence may be removed from the integral. If l is the length of the cylindrical section and q_l is the charge on the cylinder per unit length

$$\oint \mathbf{E} \cdot \mathbf{ds} = E_r \oint ds = E_r 2\pi r l = q_l \frac{l}{\kappa_0}$$

or

$$E_r = \frac{q_l}{2\pi r \kappa_0} \qquad \text{(cylinder)} \qquad (1.17)$$

These expressions for the electric field in terms of the charge or charge density are very important as these particular symmetrical charge distributions are frequently encountered in practice. As an exercise Eqs. (1.16) and (1.17) may be obtained directly from Eq. (1.15).

1.5. Capacity and Condensers.—It is frequently a great convenience to know the relation between the charges carried by a group of conductors and the potentials of the conductors. If this relation is known, a measurement of the potentials will yield the charges and vice versa. The potential of any point, in particular a point on the surface of one of the conductors, depends on the magnitudes and positions of all the charges present. From Eq. (1.9) the relation between the potential and the charges is a linear one and hence if V_i and q_i represent the poten-

tial and charge of the ith conductor of a group of n conductors the relations between the V's and the q_i's can be written

$$
\begin{aligned}
V_1 &= p_{11}q_1 + p_{12}q_2 + p_{13}q_3 + \cdots + p_{1n}q_n \\
V_2 &= p_{21}q_1 + p_{22}q_2 + p_{23}q_3 + \cdots + p_{2n}q_n \\
V_3 &= p_{31}q_1 + p_{32}q_2 + p_{33}q_3 + \cdots + p_{3n}q_n \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
V_n &= p_{n1}q_1 + p_{n2}q_2 + p_{n3}q_3 + \cdots + p_{nn}q_n
\end{aligned}
$$

or, more briefly, n equations of the form

$$V_i = \sum_{j=1}^{j=n} p_{ij}q_j \tag{1.18}$$

The p_{ij}'s are purely geometrical coefficients known as *coefficients of potential.* They are not all independent of one another, as can be seen as follows: Assume an equilibrium distribution of charges on each conductor, and write the potential of, say, conductor j in terms of the surface distribution of charge q_{si} over each of the conductors where i runs from 1 to n including j.

$$V_j = \frac{1}{4\pi\kappa_0} \sum_i \int_{s_i} \frac{q_{si}\, ds_i}{r_{ij}}$$

where r_{ij} is the distance from the surface element ds_i to some arbitrary point on conductor j. Now assume any other choice of charges on the same conductors, say q'_i on conductor i, and multiply the previous potential V_j by q'_j writing the latter as $\int_{s_j} q'_{sj}\, ds_j$ on the right. Then

$$q'_j V_j = \frac{1}{4\pi\kappa_0} \sum_i \int_{s_i} \int_{s_j} \frac{q_{si} q'_{sj}\, ds_j}{r_{ij}}$$

since the integrations are independent. If this expression is summed over j, it is seen to be quite symmetrical in the primed and unprimed quantities representing the two arbitrarily chosen sets of charges and their associated potentials.

$$\sum_j q'_j V_j = \frac{1}{4\pi\kappa_0} \sum_i \sum_j \int_{s_i} \int_{s_j} \frac{q_{si} q'_{sj}}{r_{ij}} ds_i\, ds_j = \sum_i q_i V'_i$$

This is known as *Green's reciprocation theorem.* The condition on the p_{ij}'s can be immediately derived from this theorem. Assume that the first distribution of charge is that for which all the q_i's are zero except q_k; then $V_i = p_{ik}q_k$. Assume the second distribution of charge is that

for which all the q_i''s vanish except q_l'; then $V_j' = p_{jl}q_l'$. On multiplying the first equations by the appropriate q_i''s, recalling that these are all zero except q_l', and equating the sum to the sum of products of the second equations by the appropriate q_j's, recalling that all these vanish except q_k, the result is

$$V_l q_l' = p_{lk} q_k q_l' = V_k' q_k = p_{kl} q_l' q_k$$

or

$$p_{lk} = p_{kl}$$

Equation (1.18) can be solved explicitly for the q's in terms of the

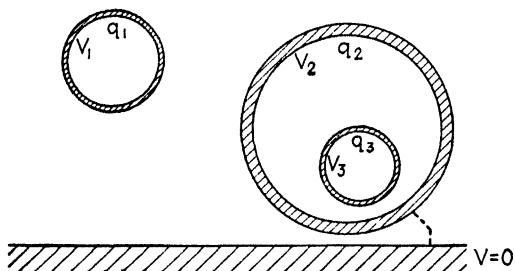



FIG. 1.15.—A system of three conductors and earth. Conductor 2 completely surrounds conductor 3.

V's. This can be done most conveniently by means of determinants.[1] The equation then becomes

$$q_i = \sum_{j=1}^{j=n} c_{ij} V_j \tag{1.19}$$

where $c_{ij} = A_{ij}/D$. Here D is the determinant of the coefficients of the q's in Eq. (1.18) and A_{ij} is the cofactor of p_{ij} in that determinant. The equality of c_{ij} and c_{ji} follows from that of p_{ij} and p_{ji}. The c_{ij}'s are known as *coefficients of induction* and the c_{jj}'s as *coefficients of capacity*. Since a positive charge on a single conductor gives it a positive potential, the coefficients of capacity are inherently positive. Consider that there are only two conductors, one of which is connected to the earth, which may be thought of for these experiments as an infinite reservoir of charge at an arbitrary potential zero. If the isolated conductor has a positive charge and hence a positive potential, a negative charge will be induced on the other conductor. Hence the coefficients of induction are in general negative, though in some cases they are zero.

As an example consider the three conductors shown in Fig. 1.15. Assume first that $q_3 = 0$. In this case the discussion of the preceding section shows that $V_3 = V_2$. If conductor 2 is connected to the earth,

[1] See any algebraic text.

which is assumed to be at the potential zero, the three equations for these three conductors become

$$q_1 = c_{11}V_1 \qquad q_2 = c_{12}V_1 \qquad 0 = c_{13}V_1$$

From these it is evident that c_{13} vanishes. Thus the coefficient of induction between two conductors that are completely shielded from one another by a conducting surface is zero. If q_3 is not zero but conductor 1 and conductor 2 are both connected to the earth so that

$$V_1 = V_2 = 0$$

the three equations become

$$q_1 = 0 \qquad q_2 = c_{23}V_3 \qquad q_3 = c_{33}V_3$$

Since there can be no net charge inside a Gaussian surface lying in conductor 2 and surrounding the cavity, q_2 lying on the inside walls of the cavity, must equal $-q_3$. Therefore $c_{23} = -c_{33}$ and the coefficient of induction between two conductors, one of which completely surrounds the other, is equal in magnitude but opposite in sign to the coefficient of capacity of the inner conductor. Two conducting surfaces such as 2 and 3 of Fig. 1.15 are said to form a *condenser*. The capacity of the condenser is the coefficient of capacity of the inner conductor. If one conductor is merely in the presence of others, its coefficient of capacity is frequently referred to as a *stray capacity*. The unit in which capacity is measured is evidently the coulomb per volt. Since the volt is the same as the joule per coulomb, the unit of capacity is the coulomb squared per joule. The name *farad* is given to the unit of these dimensions. A condenser has the capacity of 1 farad if equal and opposite charges of 1 coulomb on the two surfaces produce a difference of potential of 1 volt between them.

The conventional method of representing a condenser in a diagram is shown in Fig. 1.16. The heavy lines represent the two conducting surfaces of the condenser with charges q and $-q$. The difference of potential of the plates is V and by definition the capacity C is the ratio q/V. Condensers are frequently used in combination. The two arrangements most frequently encountered are also shown in Fig. 1.16. In the series type of circuit each plate of one condenser is connected to the adjacent plate of the next. Assume that the condensers are originally uncharged and that a charge q is placed on one of the terminating plates of the series. Equal and opposite charges are then induced on the plates of all the rest. The potential differences across the plates of the condensers are then

$$V_1 = \frac{q}{C_1}, \qquad V_2 = \frac{q}{C_2}, \qquad V_3 = \frac{q}{C_3}$$

The total potential difference between the terminating plates is the sum of these individual potential differences or

$$V = V_1 + V_2 + V_3 = \frac{q}{C}$$

where

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \qquad \text{(series)} \tag{1.20}$$

C is the ratio of the charge on the terminating plates to the difference in potential between them and hence is the effective capacity of the

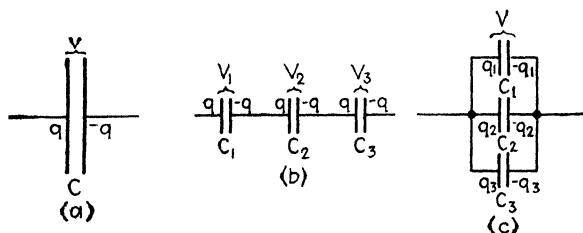



FIG. 1.16.—Combinations of condensers. (a) Condenser. (b) Condensers in series. (c) Condensers in parallel.

combination. The capacity of condensers in series is thus seen to be the reciprocal of the sum of the reciprocals of the individual capacities. The other common connection is the parallel one in which one plate of each condenser goes to a common terminal. The potential difference across each is, say, V and the charges on the individual condensers are $q_1 = C_1V$, $q_2 = C_2V$, $q_3 = C_3V$. Therefore the total charge is

$$q = q_1 + q_2 + q_3 = CV$$

where

$$C = C_1 + C_2 + C_3 \qquad \text{(parallel)} \tag{1.21}$$

Therefore the ratio of the total charge to the potential difference for this combination, which is the effective capacity, is the sum of the capacities of the individual condensers.

The capacity of a condenser can be readily calculated for certain simple geometries. Consider the concentric spherical surfaces of Fig. 1.17. The electric field in the region between the spheres is given by Eq. (1.16). The integral of E_r from b to a gives the potential difference between the surfaces. If the outer sphere is grounded ($V_b = 0$) and V is written for V_a

$$V = -\int_b^a E_r\,dr = \frac{q}{4\pi\kappa_0}\left(\frac{1}{r}\right)_b^a = \frac{q}{4\pi\kappa_0}\left(\frac{1}{a} - \frac{1}{b}\right)$$

Or as the coefficient of q is the reciprocal of the capacity C

$$C = 4\pi\kappa_0 \frac{ab}{b - a} \qquad \text{(spherical condenser)} \qquad (1.22)$$

If the radius of the outer sphere is assumed to become very large, the inner sphere becomes approximately an isolated sphere in free space. In this case the capacity is seen to approach the value

$$C = 4\pi\kappa_0 a$$

Therefore the capacity of an isolated sphere is proportional to its radius.

The capacity per unit length of a condenser formed by two coaxial

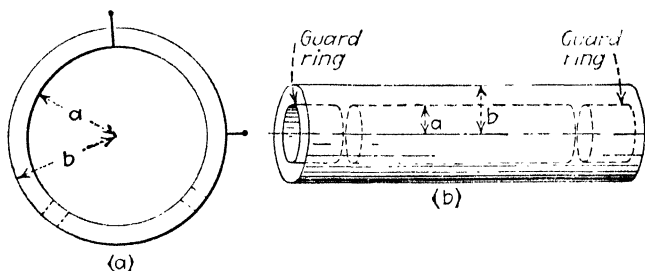



FIG. 1.17.—(a) Spherical condenser. (b) Cylindrical condenser with guard rings.

cylindrical surfaces can be calculated in a similar way. Equation (1.17) is the expression for the field outside an infinite cylinder. For it to apply to a cylinder of finite length the effects of the ends must be negligible. A method of minimizing the effects of the ends is illustrated in Fig. 1.17. The inner cylinder is divided into three sections. The central section and the outer surface form the condenser proper and the two terminating sections are known as *guard rings*. The sections are insulated from one another and from the outer cylinder, but the central section and guard rings are at the same potential. If the separation between the sections and the difference in radius of the cylinders are small compared to the length of the terminating sections, the field in the neighborhood of the central cylinder is radial to a close approximation and Eq. (1.17) applies. Integrating the field from the outer cylinder of potential 0 to the inner cylinder of potential V

$$V = -\int_b^a E_r\,dr = \frac{-q_l}{2\pi\kappa_0}\int_b^a \frac{dr}{r} = \frac{-q_l}{2\pi\kappa_0}(\log_e r)_b^a = \frac{q_l}{2\pi\kappa_0}\log_e\frac{b}{a}$$

Or the capacity per unit length, C_l, which is the reciprocal of the coefficient of q_l, is

$$C_l = \frac{2\pi\kappa_0}{\log_e (b/a)} \qquad \text{(cylindrical condenser)} \qquad (1.23)$$

If the inner section has a length l, the capacity of this section is $C_l l$.

The capacity per unit area of two plane parallel plates of infinite extent can be obtained immediately from Eq. (1.15). By symmetry the field must be uniform and equal to the value E_n in the region between the plates. If the separation of the plates is d, the difference in potential between them is simple E_nd or $\frac{dq_s}{\kappa_0}$. Therefore the capacity per unit area is

$$C_s = \frac{\kappa_0}{d} \qquad \text{(parallel-plate condenser)} \tag{1.24}$$

This expression is accurate if the assumption of a uniform field is fulfilled. Thus the separation between the plates must be small in comparison

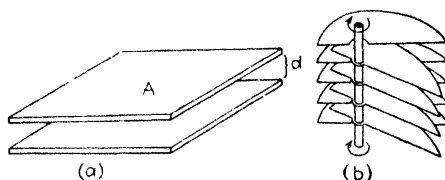



FIG. 1.18.—(*a*) Parallel-plate condenser, $C = \frac{\kappa_0 A}{d}$. (*b*) Rotary variable type of parallel-plate condenser.

with their linear dimensions or a guard rim must be provided so that the field is uniform over the portion of the plates under consideration. If the approximation is valid and the area used is A, the capacity is C_sA or

$$C = \frac{\kappa_0 A}{d} \tag{1.24'}$$

The parallel-plate condenser is the form that is most widely used in all types of electrical work. By making A large and d small C can be made very large and the pair of surfaces can be used for the storage of large quantities of electrical charge without involving excessive differences of potential.

A parallel-plate condenser of variable capacity can be constructed by arranging to vary the overlapping area of the surfaces. The simplest type is the rotary variable condenser illustrated in Fig. 1.18. A stack of approximately semicircular plates are affixed perpendicular to an axis passing through their diameters. When the axis is rotated, these plates pass into a fixed stack of plates with a similar spacing. If d is the separation between one of the movable plates and an adjacent fixed plate and n is the number of movable plates, the capacity is given approximately by $2n\kappa_0 A/d$, where A is the overlapping area of the two stacks. If the movable plates are semicircles and the axis of rotation passes through their centers, $A = \theta r^2/2$ where r is the constant radius of the plates and θ is the angle through which the movable plates have entered the fixed ones. Therefore, since $C = n(\kappa_0 r^2/d)\theta$ the capacity is pro-

portional to the angle of rotation. For various special purposes it is frequently desirable to have the capacity vary in some other manner as the plates are rotated. As an example, when dealing with resonant circuits (Sec. 13.2), a linear dependence of resonant frequency on angle of rotation of the condenser is often desired. This may be accomplished by shaping the rim of the plates so that the radius is a particular function of the azimuth angle. For a linear dependence of frequency on θ, $1/\sqrt{C}$ must vary linearly with θ or $\theta^2C = \text{const.}$ Therefore $\frac{dC}{d\theta} = \frac{-2C}{\theta} = \frac{\text{const.}}{\theta^3}$. But since the change in capacity for a small rotation is proportional to the change in area that takes place in the region of entry where the radius of the plates is, say, r, $dC = \text{const.}\ r\ r\ d\theta$ or $dC/d\theta = \text{const.}\ r^2$. Equating these values of $dC/d\theta$, the equation of the rim of the plates is seen to be $r^2\theta^3 = \text{const.}$ This is known as a *straight-line-frequency* condenser.

1.6. Measurement of the Elementary Electronic Charge.—It was stated in an earlier section that electric charge is not indefinitely divisible but that it occurs in small discrete elementary units. The correctness of this statement is attested by all the phenomena of atomic physics. The fact that charge is not continuous but occurs in units of finite magnitude is in itself a very important fact. And it is essential for any quantitative discussion of atomic phenomena to know the magnitude of this natural unit of charge. An experiment for the measurement of this quantity was devised by Millikan, and from the fact that the observations are on the motion of a minute charged drop of oil in an electric field it is known as the *Millikan oil-drop experiment.*

A spray of oil droplets is produced by an atomizer between the two horizontal surfaces of a parallel-plate condenser. A large number of these droplets are charged owing to frictional effects at the atomizer nozzle and therefore their motion is affected by the electric field existing between the condenser plates. They also tend to fall toward the earth under the influence of gravity. The region between the condenser plates is enclosed to shield the droplets as much as possible from stray and thermal convection currents. The droplets are illuminated by an arc light and one of them is selected for dark field observation through a microscope. The net force on the drop is the sum of the electric and gravitational components. If these are in opposition and the former is greater than the latter, the drop will rise. As the motion is through a viscous medium (air), the terminal velocity is quickly achieved. The relation between the terminal velocity v of a sphere of radius a moving through a medium of viscosity η and the force applied to the sphere is given by Stokes's law, if a is not too large, as

$$F = Sv$$

where

$$S = 6\pi\eta a$$

As the net force on the drop is

$$F = F_e - F_g = qE - mg$$

where q is the charge and m is the mass of the drop, the terminal velocity acquired in the field is

$$v_e = \frac{qE - mg}{S} \tag{1.25}$$

In order to avoid the necessity of knowing either m or a, a measurement of the terminal velocity of free fall under gravity is made. Let this be given by

$$v_g = \frac{mg}{S} \tag{1.26}$$

The effective mass of the drop is its volume times its effective density (density of the drop − density of the air). If ρ' is written for this

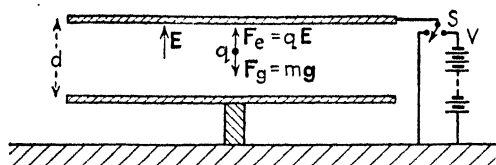



FIG. 1.19.—Principle of Millikan's experiment for measuring the electronic charge.

quantity, $m = \frac{4}{3}\pi a^3 \rho'$. Eqs. (1.25) and (1.26) are then two equations in which the only unknowns are q and a. Eliminating a

$$q = \frac{4}{3}\pi\left(\frac{9\eta}{2}\right)^{3/2}\left(\frac{v_g}{\rho'}\right)^{1/2}(v_e + v_g)\frac{V}{d}$$

where V/d is written for the electric field E.

If ions are produced by some external means in the neighborhood of the drop, the acquisition of one or more of them from time to time will change the effective charge q. Therefore the terminal velocity in the presence of the field will change. The procedure is to allow the drop to rise and fall periodically by manipulating the switch s, observing v_e during the intervals of rise and noting v_g from time to time during the periods of free fall. v_g, of course, remains constant, but it is found that v_e does not change continuously but alters by finite amounts, showing that q also can only change in finite steps. These observations also show that the alterations in q are always integral multiples of a certain unit of charge. After making an extended series of measurements that never show a change in q by an amount less than this unit it is reasonable to assume that this is the smallest unit of charge that occurs and that all charges are integral multiples of this unit. Writing e for the mag-

nitude of this unit of charge that is associated with the smallest observed change, δq, and δv_e for the corresponding change of velocity in the field

$$e = \frac{4}{3}\pi\left(\frac{9\eta}{2}\right)^{3/2}\left(\frac{v_g}{\rho}\right)^{1/2}\frac{V}{d}\,\delta v_e$$

From experiments of this type e is found to be given by[1]

$$\begin{aligned} e &= 1.60 \times 10^{-19} \text{ coulomb} \\ &= 4.80 \times 10^{-10} \text{ esu.} \end{aligned}$$

This result has been verified both directly and indirectly by other types of experiments and is of basic importance in the atomic theory of matter. In accordance with the previously adopted convention the sign of the electronic charge is shown by other experiments to be negative.

1.7. General Method of Solution of Electrostatic Problems.—If the potential is known as a function of position throughout a region, the electrostatic problem is completely solved. The electric field can then be obtained by simple differentiation as indicated by Eq. (1.13) and this gives the force on any charge located at the point if the other charges giving rise to the field remain fixed. Likewise it will be seen later that the forces on the charged conductors which themselves give rise to the field can be calculated from a knowledge of the potential or electric field. Equation (1.10) is an expression for the potential in terms of all the charges present. However, the integration is frequently difficult to perform and an alternative method for obtaining the potential is to be preferred in many cases. A differential equation that must be satisfied by the potential can be obtained by combining the general vector theorem of flux with Gauss's theorem. The flux theorem, which is derived in Appendix D, is written

$$\oint_s \mathbf{E}\cdot d\mathbf{s} = \int_v \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\right) dv$$

or more simply

$$\oint_s \mathbf{E}\cdot d\mathbf{s} = \int_v \operatorname{div} \mathbf{E}\, dv \tag{1.27}$$

where div **E** is the abbreviation for the scalar integrand in the preceding expression and is read "divergence of **E**." This may be stated in words by saying that the surface integral of the normal component of the vector **E** over a closed surface is equal to the integral of the divergence

[1] The present most precise value of e comes from electrolytic conduction experiments and a knowledge of the number of molecules per gram molecule (Avogadro's number). The accepted value [Birge, *Rev. Mod. Phys.*, **13**, 233 (1941)] is

$$e = (1.60203 \pm 0.00034) \times 10^{-19} \text{ coulomb}$$

of **E** over the volume enclosed by the surface. Writing Eq. (1.14) in the form

$$\oint_s \mathbf{E} \cdot d\mathbf{s} = \frac{1}{\kappa_0}\int_v q_v \, dv$$

where q_v is the volume charge density throughout the enclosed region, these two expressions for $\oint_s \mathbf{E} \cdot d\mathbf{s}$ may be equated, yielding

$$\int_v \operatorname{div} \mathbf{E} \, dv = \frac{1}{\kappa_0}\int_v q_v \, dv$$

Since this is true for any arbitrary volume, the integrands on the two sides must be equal at any point, for otherwise the equation would not be true for a small volume including the point. Thus

$$\operatorname{div} \mathbf{E} = \frac{q_v}{\kappa_0} \tag{1.28}$$

This states that the divergence of **E** at a point is proportional to the charge density at that point.

If **E** is eliminated between Eqs. (1.13) and (1.28), the resulting equation is

$$\operatorname{div} \mathbf{grad}\, V = -\frac{q_v}{\kappa_0}$$

By Eq. (1.11) the x component of **grad** V is $\frac{\partial V}{\partial x}$ and hence the first term in the divergence of the gradient is $\frac{\partial^2 V}{\partial x^2}$. The other terms yield analogous second partial derivatives and therefore

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = -\frac{q_v}{\kappa_0}$$

More briefly

$$\nabla^2 V = -\frac{q_v}{\kappa_0} \tag{1.29}$$

where ∇^2 is merely an abbreviation for this second partial differential operator which is called the Laplacian.[1]

The potential must be a solution of Eq. (1.29) which is known as *Poisson's equation.* If the region is free of charges and the field is produced by charged surfaces or by charges that can be excluded from the region by closed surfaces drawn about them, Poisson's equation reduces to

$$\nabla^2 V = 0 \tag{1.30}$$

[1] ∇ is read "nabla," and the Laplacian as "nabla square."

throughout the charge-free region. This is known as *Laplace's equation* and the determination of the potential in a charge-free region is reduced to the problem of finding the appropriate solution of this equation. Since Eq. (1.30) is a second-order partial differential equation, its solution for each charge-free region contains two arbitrary functions that can be evaluated either by a knowledge of the potential at all surfaces or by a knowledge of the potential and the behavior of the first derivative of the potential on a selected number of surfaces. If a solution of Eq. (1.30) can be found which satisfies the prescribed boundary conditions, it gives the value of the potential at every point of the region. The boundary conditions at the surface of a conductor are illustrated in Fig. 1.20. The potential must reduce to the potential of the conductor and the electric field which is $-\mathbf{grad}\ V$ must be normal to the surface. The methods of finding solutions of Laplace's equation that satisfy various imposed boundary conditions is the subject of potential theory and a complete discussion will be found in more mathematical treatises.[1]

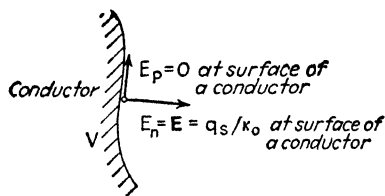



FIG. 1.20.—Boundary conditions at the surface of a conductor.

It is very useful to be able to visualize the geometrical properties associated with an electric field. Consider as an example the field inside a condenser. Let the equations of the bounding conducting surfaces at potentials V_1 and V_2 be: $F_1(x, y, z) = 0$ and $F_2(x, y, z) = 0$. Consider a point in the region between them at a potential V, where $V_1 < V < V_2$. If a test charge at the point is displaced an amount $\mathbf{dl}$, the change in potential is by Eq. (1.6) $-\mathbf{E} \cdot \mathbf{dl}$. The greatest change occurs when the vectors $\mathbf{E}$ and $\mathbf{dl}$ are parallel, *i.e.*, when the displacement is parallel to the field. This is the direction of the gradient of V ($\mathbf{grad}\ V$). The process is analogous to the motion of a mass on a sloping surface; the greatest amount of work is done when the motion is in the direction of greatest acclivity. The change in potential is zero if the displacement $\mathbf{dl}$ is perpendicular to $\mathbf{E}$. Thus an element of surface through the point perpendicular to $\mathbf{E}$ is an *equipotential surface*. This surface may be continued out from the point by following all paths that involve no work, and it will be found that the equipotential surface so described encloses the inner conductor. Any other point in the space not on this surface

[1] MAXWELL, "Electricity and Magnetism," 3d ed., Oxford University Press, New York, 1904; JEANS, "Electricity and Magnetism," 4th ed., Cambridge University Press, Cambridge, 1923; SMYTHE, "Static and Dynamic Electricity," McGraw-Hill Book Company, Inc., New York, 1939; STRATTON, "Electromagnetic Theory," McGraw-Hill Book Company, Inc., New York, 1941.

may be chosen and the equipotential passing through this point described in a similar manner. The two surfaces cannot intersect as by definition the two points are at different potentials. In this manner the entire space between the two conductors can be filled with equipotential surfaces. They are everywhere perpendicular to the electric field. A line through the space that is everywhere tangent to the field is called a *line of force.* Thus the lines of force are perpendicular to the equipotentials. Any equipotential may be replaced by an uncharged conducting shell without in any way affecting the electric field. Or the charge may be removed, say, from the inner conductor and placed on this shell and the field between it, and the outer conductor remains unchanged. The field between it and the inner conductor vanishes.

The differential equation of a line of force may be obtained from the fact that a displacement **dl** along the line is in the direction of **E**. That is, the components of **dl** must be proportional to the components of **E** or $dx/E_x = dy/E_y = dz/E_z$. A closed tubular surface made up of a series of contiguous lines of force is known as a *tube of force.* It is everywhere perpendicular to the equipotential surfaces and there is no component of **E** normal to the tubular surface. If Gauss's theorem is applied to such a tube terminated by infinitesimal areas of equipotential surfaces, it is evident that the product of the field and the area is equal for the two ends, *i.e.*, $E_1\,ds_1 = E_2\,ds_2$ as indicated in Fig. 1.21. This product of field and area is known as electric *flux.* The field strength is inversely proportional to the tube area and the tube may be considered as carrying flux from one conductor to the other in much the same manner as a fluid is transported through a pipe.

FIG. 1.21.—Equipotentials and lines and tubes of force.

1.8. Special Methods for Handling Electrostatic Problems.—The equipotentials produced by configurations of point charges are very useful in solving a number of common electrostatic problems. Figures 1.22 and 1.23 illustrate the forms of the equipotential surfaces produced by two point charges of equal magnitude. The heavy lines represent the intersections of the equipotential surfaces and the plane of the paper in which the two charges lie. The surfaces are symmetrical about the axis joining the charges, *i.e.*, if the figure is rotated about this axis, the lines will trace out the equipotential surfaces. From Eq. (1.9) the equation for the equipotentials of Fig. 1.22 is

$$q\left(\frac{1}{\sqrt{(x-d)^2+y^2+z^2}}+\frac{1}{\sqrt{(x+d)^2+y^2+z^2}}\right) = 4\pi\kappa_0 V \quad (1.31)$$

where $2d$ is the separation of the charges and V takes on successive values in passing from one curve to the next. The negative charges from which these positive charges have been separated are considered to be located at a very great distance. Consider an isolated conducting surface of the form of one of these equipotentials, say the one for which $V = V'$. The dimensions of the conductor then determine the value of d. Assume that it carries a charge $2q$. Its potential is then determined as V'. The potential at any external point is given as a function of x, y, and z by Eq. (1.31). From this equation **E** may be determined at any point

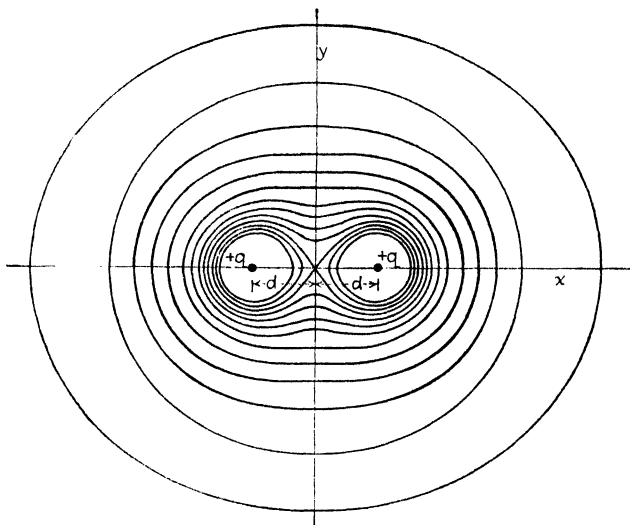



FIG. 1.22.—Equipotentials in the presence of two equal point charges (plane $z = 0$).

by differentiation and the charge density at any point on the conductor can be found by the use of Eq. (1.15).

The equipotentials due to equal charges of opposite sign are shown in Fig. 1.23. The equation for one of these curves is the same as Eq. (1.31), except that the second term is preceded by a negative sign. The inner curves are approximately circles and to the extent that the approximation is satisfactory the equation for the potential of a point due to two equal and opposite point charges may be used to solve the problem of two oppositely charged spheres. Consider two spheres of radii a and b separated by a distance $2d$, where $a << d$ and $b << d$. If the first sphere carries a charge q and the second a charge $-q$, the potentials of the spheres are given approximately by

$$V_a = \frac{q}{4\pi\kappa_0}\left(\frac{1}{a} - \frac{1}{2d}\right)$$

$$V_b = \frac{q}{4\pi\kappa_0}\left(-\frac{1}{b} + \frac{1}{2d}\right)$$

neglecting higher powers of a/d or b/d. From these expressions one can obtain for instance the capacity of the two spheres considered as a condenser

$$C = \frac{q}{V_a - V_b} = 4\pi\kappa_0\left(\frac{1}{a} + \frac{1}{b} - \frac{1}{d}\right)^{-1}$$

$$= 4\pi\kappa_0\left(\frac{1}{a} + \frac{1}{b}\right)^{-1}\left[1 + \frac{ab}{d(a+b)}\right] \qquad (1.32)$$

to this approximation. The capacity of the two spheres is greater than

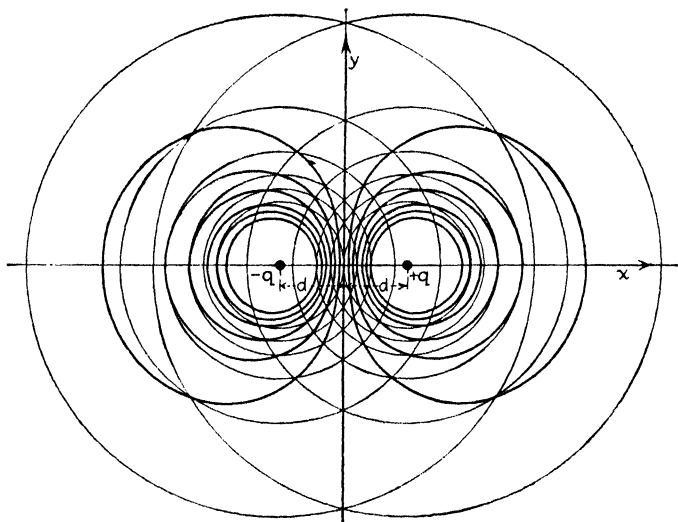



FIG. 1.23.—Heavy lines are equipotentials in the presence of two equal and opposite point charges. Light circles indicate the method of construction.

the reciprocal of the sum of the reciprocals of their separate capacities by a term of the order of a/d.

In the case of equal charges with opposite sign the central equipotential for which $V = 0$ is accurately a straight line. It represents the intersection of the diagram with an infinite plane perpendicular surface. It could be replaced by an infinite conducting plane at the potential zero (earthed) and carrying a charge $-q$ and the potential at any point in the region containing q would be unchanged. Thus the field to the right of the central vertical line produced by these two charges is the same as that produced by a point charge q a distance d in front of an infinite earthed conducting plane. The charge $-q$ which is not actually present at that position is known as an *image* by analogy with reflexion in a mirror.

The case of the point charge and plane which is depicted at the left

in Fig. 1.24 may be considered in more detail as an illustrative example. The potential of the point P in the plane of the figure ($z = 0$) having the coordinates x and y is given by

$$V = \frac{q}{4\pi\kappa_0}\{[y^2 + (x - d)^2]^{-\frac{1}{2}} - [y^2 + (x + d)^2]^{-\frac{1}{2}}\}$$

The field components are obtained by taking the negative partial derivatives with respect to x and y

$$E_x = -\frac{\partial V}{\partial x} = \frac{q}{4\pi\kappa_0}\{(x - d)[y^2 + (x - d)^2]^{-\frac{3}{2}} - (x + d)[y^2 + (x + d)^2]^{-\frac{3}{2}}\}$$

$$E_y = -\frac{\partial V}{\partial y} = \frac{q}{4\pi\kappa_0}\{y[y^2 + (x - d)^2]^{-\frac{3}{2}} - y[y^2 + (x + d)^2]^{-\frac{3}{2}}\}$$

At the plane $x = 0$, E_y is seen to vanish as it must and E_x reduces to

$$E_x = \frac{-q}{2\pi\kappa_0}d(y^2 + d^2)^{-\frac{3}{2}}$$

FIG. 1.24.—Examples of the method of images applied to point charges in front of infinite plane surfaces.

The surface density of charge on the conductor is by Eq. (1.15)

$$q_s = \frac{-q}{2\pi}d(y^2 + d^2)^{-\frac{3}{2}}$$

If this is integrated over the plane, it will be seen to reduce to $-q$, as is necessary. The force with which the charge is attracted to the plane is the same as the force of attraction between the charge and its image, or

$$F = -\frac{q^2}{16\pi\kappa_0 d^2}$$

At the right in Fig. 1.24 is indicated the system of images for the closely related problem of a charge in the acute angle formed by two semi-infinite grounded conducting planes intersecting at right angles. The horizontal pairs of charges maintain the vertical plane at the potential zero and the vertical pairs accomplish the same thing for the horizontal plane. Thus the potential produced by these four charges is the appropriate potential for any point in the acute angle.

The method for the graphical construction of equipotentials due to two charges is indicated by the light circles of Fig. 1.23. The radii of the circles about the charge q are of the form $q/n\alpha$, where α is a constant determining the scale factor and n takes on a series of integral values 1, 2, 3, etc. The potential at a point on the nth circle

due to that central charge is $V_n = n\alpha/4\pi\kappa_0$. As the charges are equal in Figs. 1.22 and 1.23 an identical series of circles is drawn about the second charge. (If the charges had been unequal, corresponding radii would have been proportional to the magnitude of the charge at the center.) The equipotentials are then drawn by connecting the points of intersection of the circles for which the sum or difference of the n's is a constant depending on whether the charges are of like or unlike sign. The lines of force corresponding to the charges in Figs. 1.22 and 1.23 are shown in the lower and upper portions, respectively, of Fig. 1.25. These lines are perpendicular to those of the previous figures. The spacing between the equipotentials of Figs. 1.22 and 1.23 and also between the lines of force in Fig. 1.25 is inversely proportional to the field strength in that region.

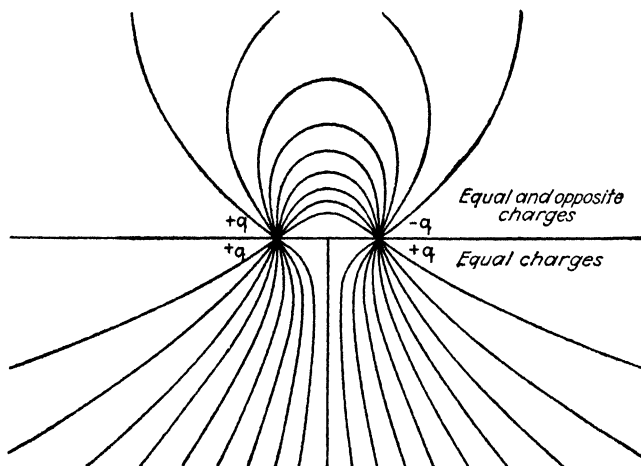



FIG. 1.25.—Lines of force produced by charges of equal magnitude.

There is one important practical case in which the two charges are of unequal magnitude and sign. Consider a charge q situated a distance d to the right of the origin and a charge of magnitude $(a/d)q$ and of negative sign a distance a^2/d to the right of the origin as shown in Fig. 1.26. The potential of a point P having the polar coordinates r and θ due to these two charges is

$$
\begin{aligned}
V_P &= \frac{1}{4\pi\kappa_0}\left(\frac{q}{r_2} - \frac{aq}{dr_1}\right) \\
&= \frac{q}{4\pi\kappa_0}\left[(d^2 + r^2 - 2rd\cos\theta)^{-\frac{1}{2}} - \left(a^2 + \frac{r^2d^2}{a^2} - 2rd\cos\theta\right)^{-\frac{1}{2}}\right]
\end{aligned}
$$

This is seen to vanish if $r = a$ for any value of θ. That is, the sphere $r = a$ is at the potential zero. Therefore this spherical equipotential may be replaced by a grounded conducting sphere. The problem represented by the above potential function is that of a grounded conducting sphere in the presence of a charge q at a distance d from its center. The radial field at any point is obtained by taking the negative partial derivative with respect to r. The field at right angles to r in

the plane of the figure is, from Eq. (1.6), $E_\theta = -(1/r)(\partial V/\partial\theta)$. At the surface of the sphere E_θ vanishes as it must and E_r becomes

$$E_r = -\frac{\partial V}{\partial r} = \frac{-q}{4\pi\kappa_0}\frac{d^2 - a^2}{a(d^2 + a^2 - 2ad\cos\theta)^{3/2}}$$

The surface charge density is obtained by multiplying this by κ_0. The force of attraction between the charge and the sphere is that between the charges q and $-\frac{aq}{d}$ separated a distance $\frac{(d^2 - a^2)}{d}$, or

$$F = -\frac{daq^2}{4\pi\kappa_0(d^2 - a^2)^2}$$

If d is large compared to a, the force of attraction varies as the inverse third power of the separation.

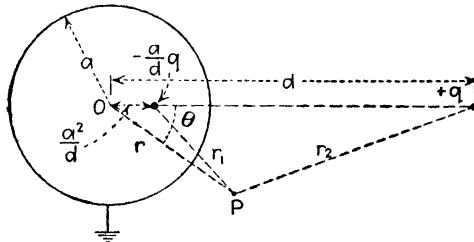



FIG. 1.26.—Method of images applied to a grounded sphere in the presence of a point charge.

If the sphere were uncharged and insulated instead of grounded, the problem would be similar except that an additional charge $+\frac{aq}{d}$ would have to be placed at the center of the sphere in order that there should be no net charge upon it. The potential would then be the sum of the potentials due to the three charges, and the sphere would remain an equipotential under the influence of the additional charge at its center.

Two-dimensional Problems.—The problem of finding the potential function corresponding to a system of conductors of effectively infinite extent along one axis and uniform cross section is somewhat simpler than the general case inasmuch as V is independent of one of the coordinates. If z is the coordinate axis parallel to the conductors, $\partial^2 V/\partial z^2$ vanishes and Eq. (1.30) becomes

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0 \tag{1.33}$$

Now it may be shown that any function of the complex variable[1]

$$z = x + jy$$

[1] Not to be confused with the third Cartesian coordinate.

(see Appendix C) is a solution of this equation. For consider

$$F(z) = F(x + jy)$$

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial z}\frac{\partial z}{\partial x} = \frac{\partial F}{\partial z} \quad \text{and} \quad \frac{\partial^2 F}{\partial x^2} = \frac{\partial^2 F}{\partial z^2}$$

$$\frac{\partial F}{\partial y} = \frac{\partial F}{\partial z}\frac{\partial z}{\partial y} = j\frac{\partial F}{\partial z} \quad \text{and} \quad \frac{\partial^2 F}{\partial y^2} = -\frac{\partial^2 F}{\partial z^2} \tag{1.34}$$

Therefore $\dfrac{\partial^2 F}{\partial x^2} + \dfrac{\partial^2 F}{\partial y^2} = 0$ and F is a solution of Eq. (1.33). Furthermore, if F is separated into its real and imaginary parts and written

$$F(x + jy) = U(x,y) + jV(x,y)$$

both U and V must be solutions of Laplace's equation. But from Eqs. (1.34)

$$\frac{\partial F}{\partial y} = j\frac{\partial F}{\partial x}$$

Therefore

$$\frac{\partial U}{\partial y} + j\frac{\partial V}{\partial y} = j\left(\frac{\partial U}{\partial x} + j\frac{\partial V}{\partial x}\right)$$

and equating real and imaginary parts:

$$\frac{\partial U}{\partial x} = \frac{\partial V}{\partial y} \quad \text{and} \quad \frac{\partial U}{\partial y} = -\frac{\partial V}{\partial x} \tag{1.35}$$

But Eqs. (1.35) are the equations that express the fact that the surfaces $U(x,y) =$ const. and $V(x,y) =$ const. meet everywhere perpendicularly. Therefore, if one of the components of the function F, say U, when set equal to a constant, represents an equipotential surface, the equation $V =$ const. contains the associated lines of force. This method of finding solutions U and V for Laplace's equation in two dimensions is a very important one. The physical problems for which these are the appropriate potential functions are evident from plots of the families of curves $U(x,y) =$ const. and $V(x,y) =$ const.

The systematic derivation of the appropriate $F(z)$ for the problem to be solved is a matter of some complexity, and the reader is referred to the more mathematical treatises mentioned earlier for a general discussion. However, it is easy to plot the results of choosing simple functions, and these generate many interesting and practical cases. The derivative of the function with respect to z gives immediately certain basic information.

$$\frac{dF}{dz} = \frac{(\partial U/\partial x)\,dx + (\partial U/\partial y)\,dy + j\,(\partial V/\partial x)\,dx + j\,(\partial V/\partial y)\,dy}{dx + j\,dy}$$

$$= \frac{\partial U}{\partial x} - j\frac{\partial U}{\partial y} = \frac{\partial V}{\partial y} + j\frac{\partial V}{\partial x}$$

where Eqs. (1.35) have been used. Or from Eqs. (1.13)

$$\frac{dF}{dz} = (E_x^2 + E_y^2)^{1/2}e^{j\theta}, \qquad \theta = \tan^{-1}(-E_y/E_x),\ U = \text{potential function}$$
$$\theta = \tan^{-1}(E_x/E_y),\ V = \text{potential function}$$

Thus the absolute magnitude of dF/dz, $|dF/dz|$ gives the field strength in either case, and the vector is directed along the lines of force, perpen-

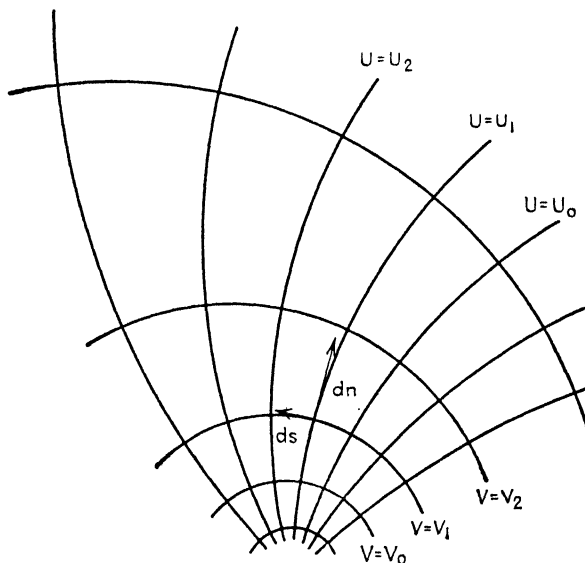



FIG. 1.27.—Infinitessimal vectors **dn** and **ds** parallel and normal, respectively, to the orthogonal surfaces U = const. and V = const.

dicular to the surfaces U = const. or V = const. whichever are chosen as equipotentials. If V is chosen as the potential function and $V(x,y) = V_1$ = const. is the equation of a conducting surface, the charge density at any point on the surface is given by Eq. (1.15) as $\kappa_0|dF/dz|$. The integral of this quantity along the conductor between two points yields the charge per unit length perpendicular to the figure between these two points. If dn is an infinitesimal vector normal to $V(x,y) = V_1$ and ds is a vector parallel to this curve (normal to the surfaces U = const.),

$$\left|\frac{dF}{dz}\right| = -\frac{\partial V}{\partial n} = +\frac{\partial U}{\partial s}$$

and along the curve $V(x,y) = V_1$

$$\int_{U_1}^{U_2} \kappa_0\left|\frac{dF}{dx}\right| ds = \kappa_0\int_{U_1}^{U_2}\frac{\partial U}{\partial s}\,ds = \kappa_0(U_2 - U_1)$$

Assuming the curve to be closed, the integral completely around it vanishes if the surface it represents is uncharged. If it and an enclosing curve represent sections through the surfaces of a condenser, the capacity per unit length of the condenser so formed is clearly

$$C_l = \frac{\kappa_0 \oint_{U_1} |dF/dz|\, dz}{|V_1 - V_2|} = \frac{\kappa_0 (U)}{|V_1 - V_2|}$$

where (U) stands for the integral of dU completely around the inner conductor and $|V_1 - V_2|$ is the potential difference between inner and outer conductor.[1]

The simplest example is that of $z = x + jy$. In this case the families of curves are $x =$ const. and $y =$ const. which represent straight lines parallel to the coordinate axes. These are the equipotentials and lines of force between the surfaces of an infinite parallel-plate condenser. The appropriate function for the cylindrical condenser is $\log_e (z)$. Writing the complex variable z in polar coordinates $z = re^{j\theta}$ (Appendix C), the function may be written immediately in terms of its real and imaginary parts:

$$\log_e z = \log_e r + j\theta.$$

$\log_e r =$ const. represents one of the family of coaxial equipotentials of the cylindrical condenser and $\theta =$ const. is a radial line of force. Writing r for the radius of the equipotential $V_r =$ const. and a for the radius of the inner cylinder, the discussion of the cylindrical condenser shows that

$$V_a - V_r = \frac{-q_l}{2\pi\kappa_0} \log_e a + \frac{q_l}{2\pi\kappa_0} \log_e r$$

As this is to be true for any value of r

$$V = \frac{-q_l}{2\pi\kappa_0} \log_e r \tag{1.36}$$

This determines the numerical value of V in terms of r and q_l, the charge per unit length on the inner cylinder. In order not to exclude the region within this cylinder, it may be contracted to an infinitesimal radius and considered as a line charge of linear density q_l.*

The potential produced by two line charges of density q_l and $-q_l$ situated a distance d to the right and left of the origin, respectively may be written

$$V = \frac{-q_l}{2\pi\kappa_0} \log_e r_1 + \frac{q_l}{2\pi\kappa_0} \log_e r_2 = \frac{q_l}{2\pi\kappa_0} \log_e \frac{r_2}{r_1} \tag{1.37}$$

[1] The numerator may be conveniently evaluated by Cauchy's integral theorm.

* The zero of potential evidently occurs at $r = 1$.

Here r_1 and r_2 are the distances from the positive and negative charges, respectively, to the point in question. From this equation it is evident that the appropriate complex function is $\log_e [(r_2 e^{j\theta_2})/(r_1 e^{j\theta_1})]$ and therefore the lines of force are given by the imaginary part as the curves

$$\theta_2 - \theta_1 = \text{const.}$$

These are the circles passing through the two points $x = d$ and $x = -d$ shown dashed in Fig. 1.28. Considering V as a parameter that can take on a series of values, Eq. (1.37) also represents a family of circles.

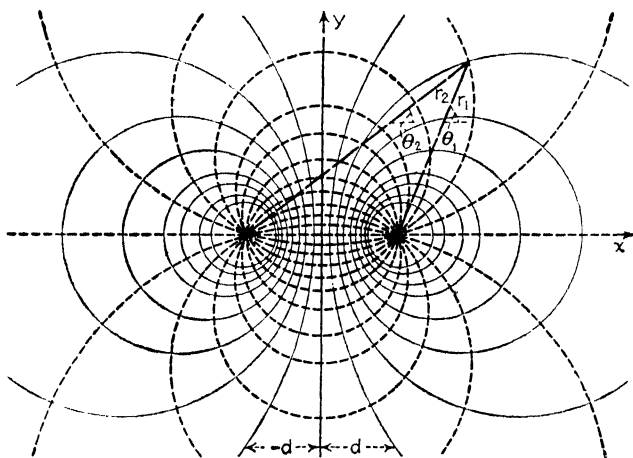



FIG. 1.28.—Families of orthogonal circles for the discussion of problems involving circular cylindrical conductors of infinite length.

These are the solid circles of Fig. 1.28. Since $r_2^2 = (x + d)^2 + y^2$ and $r_1^2 = (x - d)^2 + y^2$, Eq. (1.37) may be written

$$e^{2V'} = \frac{r_2^2}{r_1^2} = \frac{(x + d)^2 + y^2}{(x - d)^2 + y^2}$$

where V' is written for $2\pi\kappa_0 V/q_l$. Simplifying this expression

$$y^2 + \left(x - \frac{e^{V'} + e^{-V'}}{e^{V'} - e^{-V'}}d\right)^2 = \left(\frac{2}{e^{V'} - e^{-V'}}\right)^2 d^2$$

The coefficient of d on the left is the hyperbolic cotangent of V' (coth V') and the coefficient of d^2 on the right is the square of the hyperbolic cosecant of V' (csch V'). Therefore this may be written more briefly as

$$y^2 + (x - d \coth V')^2 = (d \operatorname{csch} V')^2 \tag{1.38}$$

Equation (1.38) is the equation of a circle of radius $d \operatorname{csch} V'$ with its center displaced a distance $d \coth V'$ along the x axis.

The three typical problems to which these cylindrical equipotentials

may be applied are shown in Fig. 1.29. At (*b*) in the figure are two non-coaxial cylinders of radii a and b with their axes separated a distance c. If V'_a and V'_b refer to the respective cylinders

$$a = d \operatorname{csch} V'_a \qquad \text{and} \qquad b = d \operatorname{csch} V'_b$$

The separation of the axes of the cylinders is

$$c = d(\coth V'_a - \coth V'_b)$$

These three equations may be solved for d, V'_a, and V'_b. When these quantities are known, the equipotentials and lines of force can be drawn immediately. The capacity per unit length between the cylinders is

$$C_l = \frac{q_l}{V_a - V_b} = \frac{2\pi\kappa_0}{V'_a - V'_b} \tag{1.39}$$

The force per unit length between the cylinders when they carry charges q_l and $-q_l$ per unit length, respectively, is the same as that between the

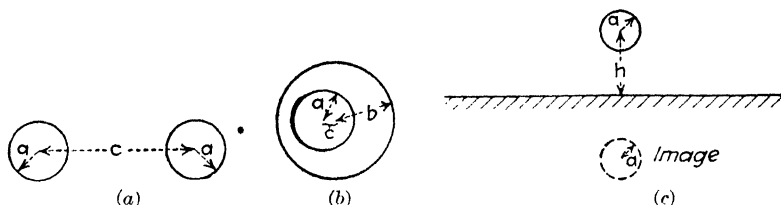



FIG. 1.29.—(*a*) Parallel conducting cylinders. (*b*) One cylindrical conductor inside another. (*c*) A conducting cylinder above an infinite plane conductor.

two line charges separated a distance $2d$. This is obtained by taking the negative partial derivative of V in Eq. (1.36) with respect to r. Setting $r = 2d$, the field at one charge due to the other is $q_l/4\pi\kappa_0 d$ and hence the force per unit length between them is

$$\frac{q_l^2}{4\pi\kappa_0 d} = \frac{C_l^2(V_a - V_b)^2}{4\pi\kappa_0 d}$$

At (*a*) in Fig. 1.29 are shown the cross sections of two conducting cylinders of equal radius. In this case $V'_b = -V'_a$ and the previous equations become

$$a = d \operatorname{csch} V'_a \qquad c = 2d \coth V'_a$$

These may be solved explicitly for V'_a and d to yield

$$d = \left[\left(\frac{c}{2}\right)^2 - a^2\right]^{1/2} \qquad \text{and} \qquad V'_a = \cosh^{-1}\left(\frac{c}{2a}\right)$$

By Eq. (1.39) the capacity per unit length is[1]

$$C_l = \frac{\pi\kappa_0}{\cosh^{-1}(c/2a)} \tag{1.40}$$

[1] Reduces to $C_l = \dfrac{\pi\kappa_0}{\log_e (c/a)}$ when $a << c$.

Diagram (c) of Fig. (1.29) represents the physical situation of a cylinder of radius a a distance h above a conducting plane. The plane corresponds to a cylinder of infinite radius. If b becomes very large, V_b' approaches zero. V_b' is therefore zero in the equation for the capacity between the cylinder and the plane. The equipotentials and lines of force above the plane are the same as in the preceding case and $h = c/2$. Therefore

$$V_a' = \cosh^{-1}\frac{h}{a}$$

$V_a' = \cosh^{-1}\frac{h}{a}$.

and

$$C_l = \frac{2\pi\kappa_0}{\cosh^{-1}(h/a)}$$

$C_l = \frac{2\pi k_0}{\cosh^{-1} h/a}$

The force of attraction per unit length between the cylinder and plane is

$$F_l = \frac{\pi\kappa_0 V_a^2}{[\cosh^{-1}(h/a)]^2(h^2 - a^2)^{1/2}}$$

These equations would apply for instance to a single telegraph line of radius a supported on poles of height h above the surface of the earth.

Problems

Problems

1. The plate of an electrophorus acquires a charge q. When the plate is brought in contact with an uncharged insulated conductor, a fraction f of this charge is transferred to it. Show that the maximum charge that can be acquired by this conductor after repeated contacts with the charged electrophorus plate is

$$\frac{fq}{(1 - f)}$$

2. Two particles each of mass m and charge q are suspended by two strings of length a from the same point. Show that the inclination, θ, of each string to the vertical is given by

$$16\pi\kappa_0 mga^2 \sin^3\theta = q^2\cos\theta$$

3. A small ball is suspended by an insulating spring at a distance d above a horizontal conducting plane. Show that a charge $q = 4(d - a)\sqrt{ka\pi\kappa_0}$ must be placed upon the ball in order to lower it a small distance a if the restoring force per unit displacement of the spring is k.

4. An infinite conducting plane at zero potential is under the influence of a point charge q. Show that the charge on any area of the plane is proportional to the angle subtended by the area at the point occupied by the charge q.

5. Show how the method of images may be used to solve the problem of a point charge situated in the acute angle π/n between two semiinfinite conducting planes where n is any integer.

6. Charge is uniformly distributed with a density q_v throughout the volume of an infinite circular cylinder of radius a (such as a uniform circular beam of electrons of

radius a). Show, by applying Gauss's theorem, that the electric field at a distance r from the axis is given by

$$E_r = \frac{rq_v}{2\kappa_0} \qquad r < a$$

$$E_r = \frac{a^2q_v}{2r\kappa_0} \qquad r > a$$

7. Show by direct integration that the electric field at a distance r from an infinite line charge of density q_l is

$$E_r = \frac{q_l}{2\pi\kappa_0 r}$$

8. A circular disk of radius R is charged to a uniform surface density q_s. Show that the electric field on the axis of the disk a distance x from the center is given by

$$E_x = \frac{q_s}{2\kappa_0}[1 - x(R^2 + x^2)^{-1/2}]$$

9. Show that it is not possible to produce a unidirectional electric field for which the magnitude of the field changes in the direction normal to that of the field.

10. It is found that the electric field toward the earth at its surface is 300 volts per meter. Show that this implies that the earth has a surface charge density of -0.00265 coulombs per kilometer^2. Consider the earth to be a conductor.

11. The electric field 1,400 m. above the earth's surface is found to be 20 volts per meter downward. Using the data of the preceding problem, show that the mean charge density in the earth's atmosphere below 1,400 m. is 1.77×10^{-12} coulomb per meter^3.

12. Show that the greatest charge that can be carried by a conducting sphere 10 cm. in radius is 3.33×10^{-6} coulomb if the surface field for which corona takes place is 3×10^6 per meter. Show that if the sphere is at a great distance from all other objects, its potential is then 3×10^5 volts. If the charge is negative, to how many electrons would this charge correspond? What would be the increase in mass if the ratio of the charge in coulombs to the mass in kilograms is 1.76×10^{11} for an electron?

13. A sphere of radius a is in the presence of an external charge q at a distance d from its center. Show that if the sphere is grounded, the ratio of the charge on the part of the sphere visible from q to that on the rest is $\sqrt{\frac{d+a}{d-a}}$.

14. Show that the force on a charge q at a distance d from an uncharged insulated conducting sphere of radius a is

$$F = -\frac{q^2a^3}{4\pi\kappa_0 d^3}\left[\frac{2d^2 - a^2}{(d^2 - a^2)^2}\right]$$

15. A charge q is placed within a metal sphere a distance r from the center. Show that there is a force attracting the charge to the inner surface of the sphere which is given by

$$F = \frac{q^2ar}{4\pi\kappa_0(a^2 - r^2)^2}$$

16. Positive charge is distributed with uniform density q_v throughout the volume of a sphere. Show that the electric field at a distance r from the center of the sphere is $E_r = q_v r/3\kappa_0$. Show that if a particle of mass m and charge $-q$ is placed in the

sphere, it will execute simple harmonic vibrations back and forth through the center with a period $2\pi(3\kappa_0 m/qq_v)^{1/2}$.

17. A soap bubble 10 cm. in radius with a wall thickness of 3.3×10^{-6} cm. is charged to a potential of 100 volts. Show that if it breaks and falls as a spherical drop, the potential of the drop is 10,000 volts.

18. Assuming a radial field of 300 volts per meter at the surface of the earth, show that a water drop 10^{-4} mm. in radius will be supported against the force of gravity anywhere above the earth's surface if it carries one electron.

19. The plates of a parallel-plate condenser are not quite parallel, the separation at one edge being $d + a$ and at the opposite one being $d - a$, where $a << d$. Neglecting edge effects show that the capacity is approximately

$$C = \frac{\kappa_0 A}{d}\left(1 + \frac{1}{3}\frac{a^2}{d^2}\right)$$

where A is the area of the plates, given that

$$\log_e\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3}\right)$$

for small x.

20. If d is the separation of the plates of a parallel-plate condenser and a piece of metal of thickness t is inserted between the plates and parallel to them, show that the capacity is increased by the amount $\kappa_0 t/[d(d - t)]$ per unit area.

21. Show that the potential produced by a line charge at a distance d from the axis of a circular conducting cylinder of radius a is that which would be produced by the actual line charge and an equal image line charge of opposite sign at a distance a^2/d toward the real charge from the center of the cylinder. What additional line charge must be assumed to be present if the cylinder is to be at zero potential? What would then be the force of attraction per unit length between the line charge and the cylinder?

22. Show that a solution of Laplace's equation in two dimensions can be written as a sum of terms of the form $r^n \sin n\theta$ and $r^n \cos n\theta$, where n can be any positive or negative integer.

23. Show that if the equation of the rim of the movable plates of a rotary variable condenser is θr^{-2} = const., the change in $\sqrt{C}$ is proportional to the change in angle of rotation. This is called a straight-line-wave-length condenser.

24. What must be the equation of the rim of the movable plates of a rotary variable condenser if the fractional change in capacity is to be independent of the angle of rotation?

25. Consider that a uniform field E_0 is produced by two very large equal and opposite charges a great distance apart. E_0 is then equal to $2q/4\pi\kappa_0 d^2$, where q is the magnitude of each charge and d is the distance from the region under consideration. (q/d^2 remains finite as q and d are indefinitely increased.) Show by the method of images that if a spherical conductor of radius a at the potential zero is present in this field, the potential at points external to it is given by

$$V = -E_0\left(1 - \frac{a^3}{r^3}\right)r \cos\theta$$

where r and θ are polar coordinates about the center of the sphere.

26. Show that the function

$$V = -E_0\left(1 - \frac{a^2}{r^2}\right)r \cos\theta$$

represents the potential produced by a grounded infinite circular cylindrical conductor of radius a placed with its axis normal to a field of strength E_0.

27. Show from Eq. (1.36) that the capacity per unit length between two infinite circular cylinders of radius a and with their axes parallel and a distance c apart is

$$C_l = \frac{\pi\kappa_0}{\log_e (c/a)}$$

if $c >> a$.

28. Show by the use of Eq. (1.36) that the capacity per unit length between two circular wires of radius a a distance c apart and each at a height h above the surface of the earth is given by

$$C_l = \frac{\pi\kappa_0}{\log_e \left(\dfrac{2hc}{a\sqrt{4h^2 + c^2}}\right)}$$

if $a << c$ and $a << h$.

29. If the two wires of the preceding problem are connected together, show that the capacity of each per unit length with respect to the earth is

$$C_l = \frac{2\pi\kappa_0}{\log_e \left(\dfrac{2h\sqrt{4h^2 + c^2}}{ac}\right)}$$

30. A wire of diameter b is wound in a helix of diameter d with a spacing c between the centers of successive turns. If d is very large in comparison with c so that successive turns may be considered as essentially parallel wires and if only the capacity between neighboring turns need be considered, show that the effective capacity between the ends of the wire (interturn capacity) is given by

$$\frac{\pi^2 d\kappa_0}{\cosh^{-1} (c/b)}$$

Show also that if c is only slightly greater than b, the interturn capacity of such a helix in micromicrofarads is approximately equal to its diameter in centimeters.

31. Show that if a charge q is placed between two grounded concentric spheres of radii a and c at a distance b from their common centers, the induced charges on the inner and outer spheres are

$$-\frac{a(c - b)}{b(c - a)}q \qquad \text{and} \qquad -\frac{c(b - a)}{b(c - a)}q$$

respectively. $(a < b < c.)$

32. Show that the capacity of a series of conductors with coefficients c_{ij} when connected together by fine wires is

$$\sum_i \sum_j c_{ij}$$

33. Two spheres of radii a and b are connected by a fine wire of length $l (l >> a, b)$. Show that if a charge Q is shared between them the tension in the wire is given by

$$\frac{abQ^2}{4\pi\kappa_0(a + b)^2 l^2}$$

34. An electron with an energy of 100 electron volts is projected parallel to a grounded conducting plate. If the point of projection is 1 mm. above the plate, show that the electron will strike the plate after traveling 41 m.

35. Two concentric spheres form the plates of a condenser. The radius of the outer sphere is fixed. Show that if the radius of the inner sphere is half that of the outer one, the field strength at the surface of the inner sphere is least for a given potential difference between them.

36. Two coaxial cylinders form the plates of a condenser. The radius of the outer cylinder is fixed. Show that if the radius of the inner cylinder is $\frac{1}{e}$ (e = base of natural logarithms) times the radius of the outer cylinder, the field strength at the surface of the inner cylinder is least for a given potential difference between them.

37. A charge is placed a distance x above the lower of two infinite parallel plates of separation d. Show that the charges induced on the upper and lower plates respectively are $-(x/d)q$ and $-[(d - x)/d]q$. This may be shown by Green's reciprocation theorem or by allowing a, b, and c of Prob. 31 to become very large, keeping their difference constant.

38. Two fine conducting wires are run through a long earthed circular conducting tube. They are parallel to the axis of the cylinder and lie in the same plane as this axis, equally spaced from it. Show that if the wires are given equal and opposite charges per unit length, there will be no force on them if their separation is $2(\sqrt{5} - 2)^{1/2}a$, where a is the radius of the tube.

39. Show that if a small hole is drilled through the wall of a charged conducting shell at a point where the electric field is E, then the new normal component of the field in the hole is $E/2$.

CHAPTER II

ELECTROSTATIC ENERGY AND DIELECTRICS

2.1. Energy Associated with a Configuration of Charges.—The preceding chapter was largely concerned with methods of finding the potential and electric field due to various configurations of charges. The field so determined is the force per unit charge on a test charge at the point. By definition the test charge is so small that it does not appreciably affect the positions of the other charges producing the field. In general the motion of charges or of charged or uncharged conductors in a field does affect the positions of the other charges. If one charged conductor is moved in the presence of another, the distribution of induced charges on the surfaces changes continuously as the relative position of the conductors is altered. In certain cases the method of images can be used to find the forces between extended conductors under these conditions by reducing the problems to one involving only point or line charges. Though such special methods can be used to find the mechanical forces in a few instances, the most general method of handling the problem is provided by the energy function. The work necessary to arrange a number of charges in position may be considered as the electrostatic potential energy of the configuration. If this potential energy is known as a function of the coordinates of the charges, the force on a charge in the direction of one of the Cartesian coordinates is the negative partial derivative of the energy function with respect to that coordinate. Therefore the first thing to be done is to find the energy associated with a given charge configuration.

Consider the elementary point charges q_1, q_2, q_3, q_4, etc., which are originally at a very great distance from one another. If q_1 is fixed and q_2 is brought up to within a distance r_{12}, the work done is

$$u_2 = \frac{1}{4\pi\kappa_0}\frac{q_1q_2}{r_{12}}$$

by Eq. (1.9). If q_3 is then brought to a point distant r_{13} from q_1 and r_{23} from q_2 while these latter charges are held in position, the work done is

$$u_3 = \frac{1}{4\pi\kappa_0}\left(\frac{q_1q_3}{r_{13}} + \frac{q_2q_3}{r_{23}}\right)$$

If q_4 is then brought into position as shown in Fig. 2.1, the work done is

$$u_4 = \frac{1}{4\pi\kappa_0}\left(\frac{q_1q_4}{r_{14}} + \frac{q_2q_4}{r_{24}} + \frac{q_3q_4}{r_{34}}\right)$$

By considering each term as the sum of two equal parts the sum of these energies may be written symmetrically as

$$U = u_2 + u_3 + u_4 = \frac{1}{8\pi\kappa_0}\left[\left(\frac{q_1q_2}{r_{12}} + \frac{q_1q_3}{r_{13}} + \frac{q_1q_4}{r_{14}}\right) + \left(\frac{q_2q_1}{r_{21}} + \frac{q_2q_3}{r_{23}} + \frac{q_2q_4}{r_{24}}\right) + \left(\frac{q_3q_1}{r_{31}} + \frac{q_3q_2}{r_{32}} + \frac{q_3q_4}{r_{34}}\right) + \left(\frac{q_4q_1}{r_{41}} + \frac{q_4q_2}{r_{42}} + \frac{q_4q_3}{r_{43}}\right)\right]$$

In this form it is evident that the first set of terms is one-half the product of q_1 and the potential of the point occupied by charge q_1 due to the other charges q_2, q_3, and q_4. Similarly, the other three sets of terms are equal to half the products of the charges and the potentials of the points occupied by them due to the other charges. Writing q_i for the charge at point i and V_i for the potential of this point due to all other charges, the above expression can be generalized to represent the energy of n point charges

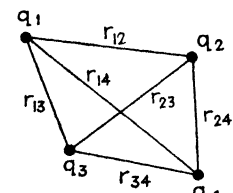



FIG. 2.1.—A configuration of point charges for discussing the associated energy.

$$U = \frac{1}{2}\sum_{i=1}^{i=n} q_iV_i \tag{2.1}$$

Equation (2.1) gives the work necessary to establish the configuration of n point charges and if the V's are known in terms of the coordinates of the charges, the forces acting on the charges can be determined by differentiation.[1]

If the charges may be considered as distributed throughout space with a volume density q_v, which is in general a function of the coordinates, Eq. (2.1) can be expressed as an integral. The charge in a volume element dv is $q_v\,dv$, and if V is the potential of the point occupied by this charge due to all the others

$$U = \tfrac{1}{2}\int q_vV\,dv \tag{2.2}$$

It is frequently convenient to have U expressed in terms of the field strength rather than in terms of the charge density and potential. Using the relation between q_v and $\mathbf{E}$ given by Eq. (1.28)

$$U = \frac{\kappa_0}{2}\int V \operatorname{div} \mathbf{E}\,dv$$

But by forming the partial derivatives of the product $V\mathbf{E}$ it is evident that

$$\operatorname{div}(V\mathbf{E}) = V\operatorname{div}\mathbf{E} + \mathbf{E}\cdot\mathbf{grad}\,V$$

[1] See Appendix A.

or

$$U = \frac{\kappa_0}{2}\int \operatorname{div}(V\mathbf{E})dv - \frac{\kappa_0}{2}\int \mathbf{E}\cdot\mathbf{grad}\, V\, dv$$

The first term may be shown in the following way to vanish if the charges are located in a finite region: By the theorem of flux (Appendix D) this volume integral may be transformed into a surface integral

$$\int \operatorname{div}(V\mathbf{E})dv = \int_s V\mathbf{E}\cdot\mathbf{ds}$$

where s is a closed surface bounding the volume containing the charges. This surface may be chosen at a very great distance from all the charges. As V decreases as $1/r$ and $\mathbf{E}$ as $1/r^2$ whereas the surface area s increases only as r^2, the integral becomes vanishingly small as $1/r$ if the surface is expanded indefinitely. Therefore the first integral can be neglected and eliminating **grad** V by means of Eq. (1.13), the energy may be written

$$U = \frac{\kappa_0}{2}\int E^2\, dv \tag{2.3}$$

where the integral is taken throughout all space. Though it is unnecessary to consider that energy, which is merely a concept, is localized in space, Eq. (2.3) is consistent with the point of view that the square of the electric intensity represents the volume density of electrostatic energy.

Since the electric field vanishes within a conductor, the energy associated with a configuration of charges is from this point of view distributed throughout empty space and in any dielectric media that may be present. It was shown by Maxwell that all the mechanical phenomena exhibited by charges and charged bodies can be accounted for by postulating an infinitely tenuous space-pervading hypothetical medium called the "ether" which is capable of transmitting stress in the manner of an elastic body. In terms of this medium there are tensions acting in the directions of the lines of force and pressures perpendicular to them. This theory was of great importance in the development and unification of the subject of electricity, and Maxwell's extension of the concept of electric current along these lines forms the basis of the theory of radiation. However, this mechanistic picture of electric forces is unnecessary for a description of the phenomena, and Eq. (2.3) may be considered as merely an alternative representation of the energy associated with a charge configuration. As an example of its use consider the mechanical force acting on the surface of a conductor. The electric field is normal to the surface and if a small area ds is assumed to be displaced outward by an amount dl, the energy $\frac{1}{2}\kappa_0 E_n^2\, ds\, dl$ will disappear from the field, for the field

vanishes within the conductor. Since the force is the negative spatial rate of change of energy, the force on the surface area ds is $\frac{1}{2}\kappa_0 E_n^2\, ds$ Thus the force per unit area acting normally to the surface of a charged conductor may be written either in terms of E_n or the surface-charge density q_s by means of Eq. (1.15) as

$$F_s = \frac{\kappa_0}{2}E_n^2 = \frac{1}{2\kappa_0}q_s^2 \tag{2.4}$$

This force determines the distortion of a flexible conducting surface when charged or when brought into an electric field. Also the integral of this force over the surface of the conductor yields the net force on the conductor as a whole.

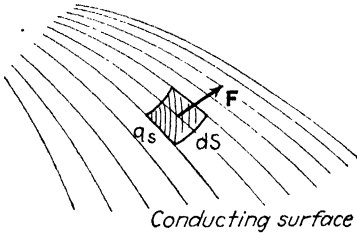



FIG. 2.2.—Mechanical force on the surface of a conductor in an electric field.

Many of the most important problems are those involving the forces which are mutually exerted on one another by a system of charged conductors. For this situation q_i and V_i of Eq. (2.1) may be considered as the charge and potential of the ith conductor of the group. Equation (2.1) may be written entirely in terms of the charges or the potentials by means of Eqs. (1.18) and (1.19)

$$\begin{aligned} U &= \frac{1}{2}\sum_{i=1}^{i=n} q_i V_i \\ &= \frac{1}{2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} p_{ij} q_i q_j \qquad (2.5) \\ &= \frac{1}{2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} c_{ij} V_i V_j \qquad (2.6) \end{aligned}$$

The geometrical factors p_{ij} and c_{ij} involve the coordinates, and if these and the charges or potentials are known, the energy is determined. The two most important cases are those in which either the charges or the potentials remain constant as the configuration is altered. The first of these can be achieved physically by insulating the conductors so that the charges on them cannot change. A small change in the energy, δU, can then be written

$$\delta U = \frac{1}{2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} \delta p_{ij} q_i q_j$$

from Eq. (2.5), where δp_{ij} is the change in this coefficient that is pro-

duced by small changes in the coordinates specifying the configuration. Considering that conductor i is moved so that its coordinate x_i is altered, the force in this direction on this conductor is

$$F_{xi} = -\frac{\partial U}{\partial x_i} = -\frac{1}{2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n}\frac{\partial p_{ij}}{\partial x_i}q_iq_j \qquad \text{(constant charge)} \qquad (2.7)$$

The second case in which the relative positions of the conductors are altered without changing their potentials requires an external source of charge and energy. For, if the positions of the conductors are changed without changing their charges, their potentials will also change. Therefore it is necessary to change the charges on the conductors by the proper amount to counteract the alteration in the potentials. The situation is realized physically by connecting the conductors to batteries which maintain their potentials at fixed values. The batteries then act as additional sources of charge and energy. From Eq. (2.1) the change in electrostatic energy associated with a change in the positions of the conductors in which the potentials remain unaltered is given by

$$\delta U = \frac{1}{2}\sum_{i=1}^{i=n}\delta q_i V_i$$

The battery, however, must supply an amount of charge δq_i to the ith conductor at the potential V_i which requires altogether an amount of energy

$$\delta U_B = \sum_{i=1}^{i=n}\delta q_i V_i$$

Thus the battery must supply twice the energy that disappears from the electrostatic field owing to the alteration. Half the energy supplied by the battery goes to the performance of mechanical work and half goes to increasing the electrostatic energy stored in the field. The change in total electrical energy is $-\delta U_B + \delta U$ or $-\delta U$, where U of course stands for the energy of the electrostatic field. Since the force in, say, the x direction on the ith conductor is minus the partial derivative of the total electrical energy with respect to the coordinate, it is plus the partial derivative of the energy in the electrostatic field alone with respect to this coordinate. As Eq. (2.6) represents the energy explicitly in terms of the potentials, it is the appropriate form to use when these are held constant and hence

$$F_{xi} = \frac{\partial U}{\partial x_i} = \frac{1}{2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n}\frac{\partial c_{ij}}{\partial x_i}V_iV_j \qquad \text{(constant potential)} \qquad (2.8)$$

Of course, the magnitudes of the force components are determined by the existing charge configuration and Eqs. (2.7) and (2.8) yield the same numerical values. The former is used when the energy is expressed in terms of the charges and the latter when it is given in terms of the potentials.

In the ordinary condenser, for instance, there are but two conducting surfaces. If the potential of one of these is chosen as zero, the energy given by Eq. (2.1) reduces to the single term $\frac{1}{2}qV$, where q is the charge on the condenser and V is the potential difference between the plates or surfaces. Since $q = VC$, or from Eqs. (2.5) and (2.6), the energy of the charged condenser can be written in the three equivalent forms

$$U = \frac{1}{2}qV = \frac{1}{2}CV^2 = \frac{1}{2}\frac{q^2}{C} \tag{2.9}$$

This is the mechanical work that must be done to charge the condenser or the energy that would be released on discharge. If the plates of the charged condenser are connected by a conducting path, the electrostatic energy is transformed into heat in the conducting path. Equation (2.4) or (2.9) may be used to find the force of attraction between the charged plates of a condenser. Consider, for example, the parallel-plate type for which the capacity per unit area is given by Eq. (1.24) as κ_0/d, where d is the plate separation. Since $E_n = V/d$, Eq. (2.4) yields

$$F_s = \frac{\kappa_0}{2}\frac{V^2}{d^2} \tag{2.10}$$

or since $\dfrac{\partial C}{\partial d} = -\dfrac{\kappa_0}{d^2}$, the same force of attraction is obtained from Eq. (2.9).

2.2. Electrostatic Instruments. *The Absolute Electrometer.*—The term "absolute" is applied to an electrical instrument when it may be used to measure electrical quantities in terms of the fundamental units of length, mass, and time. In the units here adopted these measurements are made in terms of meters, kilograms, and seconds. The absolute electrometer measures primarily the charge or potential difference in terms of these units. It may also be used to measure other electrical quantities such as capacity and rate of gain or loss of charge which is electric current. The Kelvin type of instrument is illustrated schematically in Fig. 2.3. It is essentially a parallel-plate condenser in which the central portion, which is capable of a small independent motion, is used to measure the electrostatic forces in terms of gravitational ones. The plates P_1 and P_2 represent the central portions of circular parallel plates. The lower one is supported by a quartz rod and connected by a conducting wire to the conductor whose potential relative to ground is to be measured. The upper plate is also carried by insulating supports

and the central portion of it consists of a separate circular disk of area A. This disk is supported by springs which maintain it parallel to the plates but permit small vertical displacements. The upper ends of the springs are attached to an insulating rod, by means of which the vertical position of the disk can be adjusted. The disk carries a pointer which lies between two fixed fiducial marks when the disk is exactly in the plane of P_2.

P_1 and P_2 are first connected together and a mass M is placed on the disk. The quartz rod Q_1 is adjusted till the pointer lies between the fiducial marks. Q_1 then remains fixed for the subsequent measurements and the mass M is removed. The springs then evidently raise the

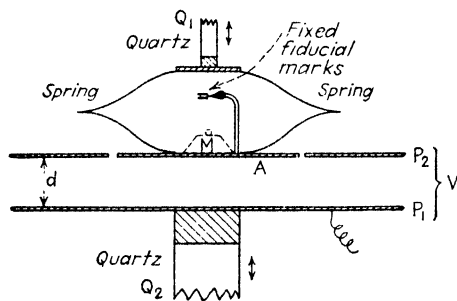



FIG. 2.3.—Absolute electrometer.

disk above the plane of P_2 and a force equal to Mg must be applied to it again before coplanarity can be restored. This force is applied electrically by establishing a difference of potential between the planes P_1 and P_2. If this potential difference is V and the plate separation is d, the electrical force on the plate of area A is evidently $\frac{1}{2}\kappa_0 V^2 A/d^2$ from Eq. (2.10). Thus the value of V that returns the pointer to its position between the fiducial marks is given by

$$Mg = \frac{\kappa_0}{2}\frac{V^2 A}{d^2} \qquad \text{or} \qquad V = \sqrt{\frac{2Mg}{\kappa_0 A}}\,d$$

All the quantities on the right are known in terms of the fundamental units of length, mass, and time, and hence V may be measured in this way in terms of these quantities. The constant κ_0 is entirely arbitrary as far as the principle of the measurement is concerned. It determines the system of electrical units in which V is measured. In the absolute practical system $\kappa_0 = 8.85 \times 10^{-12}$ farads per meter or in the electrostatic system κ_0 is unity and the units of length and mass are changed to the centimeter and gram, respectively. The sensitivity of the instrument, which may be taken as the partial derivative of the force with respect to V, is $\kappa_0 AV/d^2$. The upper limit of A and the lower limit of d are fixed by mechanical considerations. If these two factors are

constant, the sensitivity is proportional to V. Hence the instrument is most suitable for measuring large potential differences. Smaller potential differences (of the order of 10^3 volts) may be measured by applying an auxiliary unknown but constant potential V' between P_2 and ground. With P_1 grounded the position of Q_2 is adjusted till the pointer is between the marks, for which position say $d = d'$. In this condition

$$V' = \sqrt{\frac{2Mg}{\kappa_0 A}}d'$$

A small potential difference v is then applied between P_1 and ground and Q_2 adjusted for coplanarity of the suspended disk which occurs say at a plate separation d. Then $V' + v = \sqrt{\frac{2Mg}{\kappa_0 A}}d$, or, eliminating V'

$$v = \sqrt{\frac{2Mg}{\kappa_0 A}}(d - d')$$

With a well-constructed instrument the difference in height of the plate P_1, which is equal to $(d - d')$, can be read with considerable accuracy, and the small potential difference v can be directly determined.

Electrostatic Voltmeters and Electroscopes.—The simple geometry of the Kelvin electrometer permitted the calculation of the potential difference between the plates in terms of known quantities. In general, however, it is difficult to design an instrument for general utility in which the forces are calculable. In consequence most electrostatic instruments are of the type that must be calibrated by means of a previously known variable source of potential difference. This, however, presents no difficulty or objection. Familiar electrical-circuit techniques described in subsequent chapters can be used to determine the potential difference corresponding to the scale reading of an arbitrary electrostatic instrument with a high degree of accuracy. Instruments can readily be designed to measure potential differences over the range from 10^{-4} to 10^6 volts and to retain their calibration over long periods of time. These instruments are of a wide variety of types, but they all depend on the measurement of the equilibrium position of a conductor under the influence of electrical and gravitational or elastic forces.

In the higher range (above about 1 volt) these potential measuring instruments are generally known as electrostatic voltmeters. One type is similar in principle to the rotary variable condenser shown in Fig. 1.18. The movable plates, which carry a pointer or mirror for reflecting a beam of light, are held outside the fixed ones by a spiral spring or the torsion of a fiber. In equilibrium the electrical torque,

$\frac{1}{2}V^2\frac{\partial C}{\partial \theta}$ is equal to the restoring torque of the fiber after rotation through the angle θ, and may be written $\alpha\theta$ if θ is small. If α and $\partial C/\partial \theta$ can be calculated in terms of length, mass, and time, the instrument is an absolute one, but in general this is not feasible and the instrument is calibrated with a known source of potential. It is to be noted that the deflection involves the square of V and hence is in the same direction whether V is positive or negative. Thus the instrument may be used to measure alternating potentials. The chief advantage of an electrostatic instrument is that there is no change in charge at equilibrium and hence no continuous current is drawn during a measurement.

FIG. 2.4.—String electrometer or quartz-fiber electroscope.

Instruments for measuring potentials below about 1 volt are generally known as electroscopes or electrometers. Many forms of these instruments are in use, two of which are shown in Figs. 2.4 and 2.5. The simplest form of the quartz-fiber electroscope consists of a fiber covered with a thin conducting layer which is stretched halfway between two wedge-shaped electrodes. These are maintained at potentials $+V'$ and $-V'$ so that if the fiber is central and at the potential 0, it is in equilibrium. If the fiber is raised to the potential V, it acquires positive charge and is attracted toward the negative electrode. This force of attraction is counterbalanced by the distortion of the lower quartz bow and the equilibrium displacement of the fiber is a measure of its potential V. In the leaf type of electroscope a central conducting element in the form of a rigid rod carrying a hinged or flexible conducting fiber or foil is supported by an insulator within a metal chamber. If the inner element is charged, the forces due to these charges and those induced on the inner walls cause the foil to diverge from the supporting rod. The deflection of the leaf is a measure of the magnitude of the charge or of the difference of potential between the inner element and the walls. The quartz-fiber type distinguishes between the signs of charge and the leaf type does not. Both types have very small capacities, of the order of 10^{-12} farad and hence may be used to measure very minute quantities of charge. A sensitivity of 100 scale divisions per volt corresponds to 10^{-14} coulomb per division for this capacity. These instruments are used chiefly for measuring very small electric currents

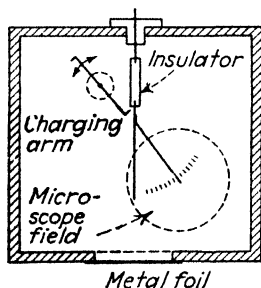



FIG. 2.5.—Leaf type of electroscope for ionization measurements.

such as those due to the ionization produced in a gas by the emanations from radioactive substances. If the electroscope of Fig. 2.5 is charged and some radioactive material placed below the metal-foil window, the leaf will fall owing to the ions of opposite sign that reach the central element. If the leaf is observed to fall at the rate of, say, 1 division in 100 sec. and the instrument has a capacity of 10^{-12} farad and a sensitivity of 100 divisions per volt, this corresponds to an acquisition of charge at the rate of 10^{-16} coulomb per second. The rate of change of charge of 1 coulomb per second constitutes an electric current of 1 *ampere;* therefore this rate of fall of the leaf will correspond to a current of 10^{-16} ampere.

The Quadrant Electrometer.—One of the most familiar instruments in electrostatic work is the quadrant electrometer illustrated schematically in Fig. 2.6. A shallow pillbox, which is generally made of brass, is cut apart along two perpendicular diameters to form a set of four quadrants. These are supported on insulating pillars, opposite ones being connected electrically. Within this structure is supported a very light plane conducting "needle" which is generally in the form of two open fans placed handle to handle. A central, light, perpendicular wire carries a mirror which is used to reflect a beam of light to a scale. This system is suspended by a fine conducting quartz fiber which supplies the mechanical restoring torque. The instrument is housed in a metal container for electrostatic shielding and provided with the necessary adjustments for leveling and altering the height and orientation of the needle.

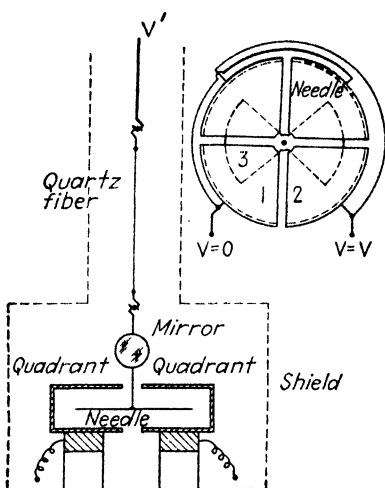



FIG. 2.6.—Quadrant electrometer.

The theory of the instrument may be developed from Eq. (2.6). Designating the two pairs of quadrants and the needle by the numbers 1, 2, and 3, respectively, and considering one pair of quadrants grounded, the other pair at the potential V, and the needle at the potential V' the energy is given by

$$U = \tfrac{1}{2}c_{22}V^2 + c_{23}VV' + \tfrac{1}{2}c_{33}V'^2$$

The c's are functions of the angular displacement θ of the needle, and to a first approximation it will be assumed that they may be written $a_{ij} + b_{ij}\theta$ for small values of θ. In equilibrium the electrostatic torque $\partial U/\partial\theta$ is balanced by the restoring torque of the fiber which may be written $k\theta$

to a sufficient approximation. The equilibrium equation is then

$$k\theta = \tfrac{1}{2}b_{22}V^2 + b_{23}VV' + \tfrac{1}{2}b_{33}V'^2$$

If the needle is placed symmetrically between the quadrants, as shown in the figure, the torque must be zero for any value of V' if $V = 0$; therefore b_{33} must vanish. In this case θ may be written

$$\theta = \frac{1}{2k}(b_{22}V + 2b_{23}V')V$$

The instrument may be used in two ways. The ideostatic connection is that in which the needle and one pair of quadrants are connected together, the other pair being grounded. In this case $V = V'$ and θ is seen to be proportional to V^2. Thus the deflection is in the same sense for either positive or negative V and the instrument may be used to measure the effective value of an alternating potential. However, the deflection is not a linear function of V, which is frequently an inconvenience, and the sensitivity is not great. In the heterostatic connection the needle is maintained at a constant high potential of the order of 100 volts, and the potential difference to be measured is applied between the quadrant pairs. In this connection the term $b_{22}V$ can generally be neglected in comparison with the other and the deflection is then proportional to $(b_{23}V'/k)V$. Thus the sign of the charge on the ungrounded quadrants is significant and the scale is linear in V. Also, as V' can be made large, the sensitivity is greater in this connection than in the other.

The scale is calibrated by applying known potential differences to the quadrants and observing the position of the light beam reflected from the mirror. The effective capacity of the instrument may be determined by means of a known standard condenser. One plate of this condenser is grounded and the other connected to a battery of potential V. If C is the capacity of the standard condenser, the charge that it acquires is $q = CV$. If the plate is then disconnected from the battery and connected to the free quadrants which were previously uncharged, the charge on the condenser and electrometer quadrants together is q. Their capacity in parallel to ground is $C + C_e$, where C_e is the effective capacity of the instrument, and if the electrometer then shows a deflection corresponding to V'

$$V' = \frac{q}{C + C_e} \qquad \text{and as} \qquad q = VC$$

$$C_e = \frac{V - V'}{V'}C$$

When C_e has been found in this way, the same technique can be employed to measure the capacity of an unknown condenser. It is evident from

the above discussion that the charge sensitivity of the instrument may be altered by placing additional capacity between the quadrants. Electric currents may be measured by recording the rate of deflection of the needle and the range may be varied by means of auxiliary capacities. Currents may also be measured by the aid of Ohm's law (Sec. 3.4). If a resistance R is placed between the free quadrants and ground, the potential of the free quadrants is iR where i is the current flowing through R. Thus, if R is known, currents may be measured by a constant-deflection method. Currents of the order of 10^{-14} amp. can be detected in this way. Smaller currents can be measured by the rate of deflection technique; the lower limit being of the order of 10^{-18} amp.

2.3. Dielectric Media.—When a piece of matter is given an electric charge, the charge may remain localized in the region of generation or application over a considerable period of time, or it may spread over the surface practically instantaneously. In the former case the substance is known as a dielectric or insulator and in the latter case as a conductor. Substances may thus be divided roughly into these two categories. The line of demarcation, however, is not sharp and there is a continuous gradation among substances from good conductors to good insulators. This matter will be discussed further in connection with the conduction of electricity, but in the present section it will be assumed for simplicity that dielectric substances are perfect insulators and will retain localized charges indefinitely. Certain substances, such as glasses, resins, and waxes, approach very closely to this ideal. Actually also the properties of a substance frequently vary from point to point, *i.e.*, it is not homogeneous. In the neighborhood of a given point the properties of the substance may not be the same in every direction, *i.e.*, it may not be isotropic. For the purposes of the immediate discussion, however, it will be assumed that the substances dealt with are perfect homogeneous isotropic dielectrics.

It was found by Cavendish, and later independently by Faraday, that the capacity of a condenser is altered if a dielectric substance is inserted between the plates. This may be shown experimentally by the use of an electrometer as described in the preceding section. If C_0 is the capacity of a condenser when the region between the surfaces is evacuated and C is its capacity when this region is filled with a dielectric, the ratio of C to C_0 is found to be independent of the shape or size of the condenser. However, this ratio is a characteristic of the particular dielectric medium that is used. In fact, this ratio is defined to be the *dielectric constant* of the substance and is written κ.

$$\kappa = \frac{C}{C_0} \tag{2.11}$$

Since the capacity is by definition the ratio of the charge to the difference

of potential between the plates, this ratio must change when the intervening space is filled with a dielectric medium. The change is small in the case of gases at atmospheric pressure, the fractional change being of the order of 0.1 per cent, but for other materials such as paraffin oil the change is by a factor of 2, and if water could be used as a perfect dielectric, the change would be by a factor of 80. The value of κ determined in this way is found to be positive for all substances, ranging from values very close to unity for gases up to the value 81.07 for pure water.

FIG. 2.7.—Effect of the intervening medium on the capacity between two surfaces. C_0 is the capacity in the absence of the medium. $C = \kappa C_0$ is the capacity with the medium present.

The physical phenomena involved may be illustrated by a consideration of a parallel-plate condenser and a slab of dielectric material, just filling the space between the plates, that may be either inserted or withdrawn. Consider first the case in which the plates are equally and oppositely charged and then insulated. When the dielectric slab is inserted between the plates, the capacity increases by the factor κ and as the charge must remain constant, this implies that the potential difference between the plates must have decreased by the factor κ. For if q_0 is the charge on the plates and V_0 and V the initial and final potential differences

$$q_0 = C_0V_0 = CV = \kappa C_0 V \qquad \text{or} \qquad V_0 = \kappa V$$

As the field is the ratio of V to the plate separation, the field in the dielectric must be smaller by the factor κ than the field that previously existed between the plates. As this is not dependent on the particular form of condenser chosen, it may be stated in general that the field due to charges placed in a region filled with dielectric material is less by the

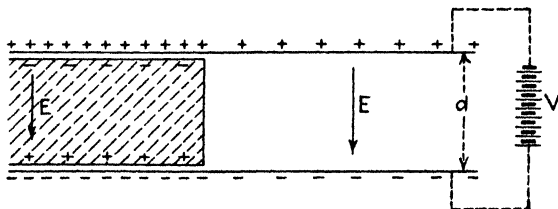



FIG. 2.8.—Dielectric slab between condenser plates.

factor κ than it would be if the medium were absent. Hence Coulomb's law for point charges in a dielectric medium would be written

$$\mathbf{F} = \frac{1}{4\pi\kappa\kappa_0}\frac{q_1q_2}{r^2}\mathbf{r}_1 \tag{2.12}$$

If, on the other hand, the plates of the condenser are connected to a battery so that the potential difference between them remains constant, the charges on the plates must increase by the factor κ when the slab is inserted. If q is the final charge on the plates

$$V_0 = \frac{q_0}{C_0} = \frac{q}{C} = \frac{q}{\kappa C_0} \qquad \text{or} \qquad q = \kappa q_0$$

Since V remains the same, the field in the region is unaltered, but the charge on the condenser plates increases. In the first case the electrostatic energy decreases by the difference between $\frac{1}{2}q_0V_0$ and $\frac{1}{2}q_0\frac{V_0}{\kappa}$. In the second case it increases by the difference between $\frac{1}{2}\kappa q_0V_0$ and $\frac{1}{2}q_0V_0$. The battery, however, supplies the energy

$$V_0(q - q_0) = V_0q_0(\kappa - 1)$$

hence an amount of mechanical work equal to $\frac{1}{2}q_0V_0(\kappa - 1)$ is performed against mechanical forces in pulling the slab between the condenser

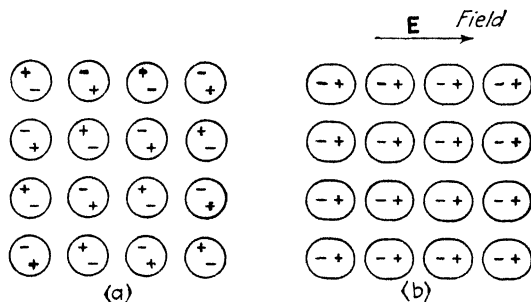



FIG. 2.9.—(*a*) Unpolarized dielectric. (*b*) Polarized dielectric.

plates. In the first case the ratio of the final to the initial electrostatic energy is $1/\kappa$ and in the second case it is κ.

2.4. General Theory of Dielectrics.—A dielectric differs from a conductor in that there are no conduction electrons in it which are free to move throughout the body of the material under the influence of an electric field. All the electrons are bound more or less tightly to the space-lattice structure of atomic nuclei, constituting a solid, or in the case of a liquid or gas the positive and negative constituents are bound together in neutral aggregates and their mass motion does not represent any net transfer of charge. However, the molecular forces holding these aggregates together are elastic in nature and if an electric field is present, the positive and negative portions tend to separate in the direction of the field resulting in both mechanical and electrical distortion. This is known as *polarization* and the result is shown schematically in Fig. 2.9. Though the charges are not free to move through the sub-

stance, a separation of charge takes place on a small scale and all the phenomena exhibited by dielectrics are attributable to the effects of polarization.

If there are equal quantities of positive and negative charge in every small region including any surfaces, the net charge density is zero and both the potential and field vanish. However, in a polarized dielectric the centers of charge are separated and this does give rise to an external field. Consider the potential due to the two charges of Fig. 2.10:

$$4\pi\kappa_0 V = q\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = q\left[\left(r^2 + \frac{a^2}{4} - ar\cos\theta\right)^{-\frac{1}{2}} - \left(r^2 + \frac{a^2}{4} + ar\cos\theta\right)^{\frac{1}{2}}\right]$$

If $a << r$, the two terms may be expanded by the binomial theorem and terms of the order of $(a/r)^2$ neglected, yielding

$$4\pi\kappa_0 V = qr^{-1}\left[1 + \frac{a}{2r}\cos\theta - 1 + \frac{a}{2r}\cos\theta\right]$$

or

$$V = \frac{qa\cos\theta}{4\pi\kappa_0 r^2} = \frac{q\mathbf{a}\cdot\mathbf{r}}{4\pi\kappa_0 r^3} = \frac{\mathbf{p}\cdot\mathbf{r}}{4\pi\kappa_0 r^3} \qquad (2.13)$$

$$\mathbf{p} = q\mathbf{a}$$
$$V = \frac{\mathbf{p}\cdot\mathbf{r}}{4\pi\kappa_0 r^3} = \frac{\mathbf{p}\cdot\mathbf{r}_1}{4\pi\kappa_0 r^2}$$

FIG. 2.10.—Potential due to an electric dipole of moment **p**.

Here **p**, which is called the electric moment of the dipole formed by the two charges, is a vector of magnitude qa in the direction from the negative to the positive charge. By taking the appropriate partial derivatives the vector **grad** $(1/r)$ is seen to be $-\frac{\mathbf{r}}{r^3}$.[1] Therefore Eq. (2.13) may be written alternatively

$$V = \frac{-1}{4\pi\kappa_0}\mathbf{p}\cdot\mathbf{grad}\,\frac{1}{r} \qquad (2.13')$$

The equipotentials which are evidently of the form $(1/r^2)\cos\theta = \text{const.}$ and the lines of force which may be shown to be the family of curves $(1/r)\sin^2\theta = \text{const.}$ are shown in Fig. 2.11. The potential due to an assemblage of dipoles is the sum of the potentials due to the individual

[1] **grad** $\left(\frac{1}{r}\right) = -\frac{1}{r^2}\left(\mathbf{i}\frac{\partial r}{\partial x} + \mathbf{j}\frac{\partial r}{\partial y} + \mathbf{k}\frac{\partial r}{\partial z}\right)$, as $r = (x^2 + y^2 + z^2)^{\frac{1}{2}}$, $\partial r/\partial x = x/r$ with analogous expressions for the other partial derivatives, and as $\mathbf{i}x + \mathbf{j}y + \mathbf{k}z = \mathbf{r}$, **grad** $\left(\frac{1}{r}\right) = -\frac{\mathbf{r}}{r^3}$.

units, and the components of the resultant electric field are obtained by taking the negative partial derivatives of this potential.

The forces acting on a dipole in a field may be obtained most conveniently from the energy point of view. Consider that the components of the dipole are first coalesced at its center and that the positive charge is given a displacement $\mathbf{a}/2$ and the negative one a displacement $-\mathbf{a}/2$.

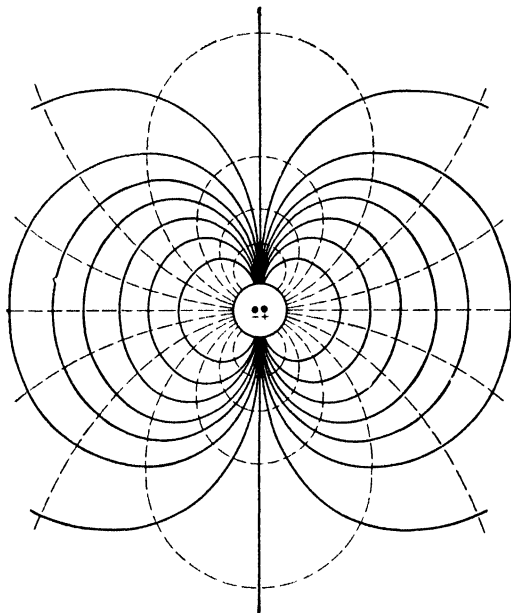



FIG 2.11.—Equipotentials, $\frac{\cos\theta}{r^2}$ = const. (solid lines), and lines of force, $\frac{\sin^2\theta}{r}$ = const. (dashed lines), associated with an electric dipole.

If no external electric field is present, work will be done against their mutual attractive forces, but this affects only the internal energy which is not under consideration.[1] The work done against the forces produced

[1] The forces that hold a physical dipole together and are responsible for its internal energy can be only partially electrostatic in nature. For it may be shown quite generally that no configuration of charges can be in stable equilibrium under purely electrostatic forces. Consider a unit positive test charge at an otherwise unoccupied point where the potential is V_0. Its increase in energy on being given a small displacement to a point of potential V is by Taylor's theorem (Appendix A):

$$V - V_0 = \left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y} + dz\frac{\partial}{\partial z}\right)(V)_0 + \frac{1}{2!}\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y} + dz\frac{\partial}{\partial z}\right)^2(V)_0$$

to second-order terms. If the force on the charge is to be zero, the first term must vanish, which is perfectly feasiable. But for the equilibrium to be stable the second term must be positive, representing an increase in energy, for any displacement. It

by an external electric field **E** is by Eq. (1.6)

$$U = -\frac{q\mathbf{a}\cdot\mathbf{E}}{2} - \frac{q\mathbf{a}\cdot\mathbf{E}}{2} = -\mathbf{p}\cdot\mathbf{E} \tag{2.14}$$

This is the energy of the previously created dipole of moment p in the electric field **E**. The forces exerted by the field on the two components of the dipole constitute a couple tending to rotate the dipole about an axis perpendicular to both **p** and **E** in such a sense that **p** and **E** tend to lie in the same direction. This torque can be obtained from Eq. (2.14) by taking the negative partial derivative of U with respect to the angular variable θ of Fig. 2.12. As $U = -pE\cos\theta$

$$T = -pE\sin\theta \tag{2.15}$$

FIG. 2.12.—Dipole of moment $\mathbf{p} = q\mathbf{a}$ in a field **E**. Energy of the dipole in the field,

$$U = -\mathbf{p}\cdot\mathbf{E}$$

Torque,

$$\mathbf{T} = \mathbf{p}\times\mathbf{E}.$$

The torque may be considered as a vector in the direction of the axis of rotation and of magnitude given by Eq. (2.15). The *vector product* of two vectors, which will be indicated by the sign $\times$, is *defined* as the vector equal in magnitude to the product of the magnitudes of the separate vectors times the sine of the angle between them and in the direction in which a right-hand screw would advance if rotated through the smaller angle in such a sense as to bring the first vector into the position previously occupied by the second (Appendix *D*). In terms of the vector product the torque on the dipole would be written

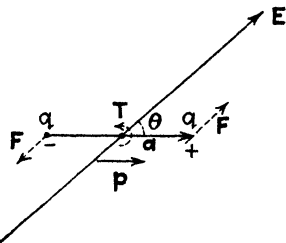
$$\mathbf{T} = \mathbf{p}\times\mathbf{E} \tag{2.15'}$$

In addition to this torque there is a net force on the dipole if the field is not constant. The components of this force are the negative partial derivatives of U with respect to the coordinates or the components of the vector $-$ **grad** U. Therefore

---

is always possible by a rotation of the coordinate system to write the second term

$$\frac{1}{2}\left(\frac{\partial^2 V}{\partial x^2}dx^2 + \frac{\partial^2 V}{\partial y^2}dy^2 + \frac{\partial^2 V}{\partial z^2}dz^2\right)_0$$

and it is evident for the particular displacement $dx = dy = dz$ that this term can be written $\frac{1}{2}\nabla^2 V\,dl^2$. Since the Laplacian vanishes at a point unoccupied by a charge a displacement can be found for which the energy does not increase, hence the equilibrium cannot be stable. In large-scale systems equilibrium may be considered to result from the combination of mechanical and electrical forces. In the case of atomic systems there must be forces which become effective at small distances that prevent the coalescing of the elementary charges. These forces have no large-scale analogues and their nature must be described by quantum mechanics, but the very existence of atomic systems composed of charged particles implies that there are forces of this type.

$$\mathbf{F} = \mathbf{grad}\,(\mathbf{p}\cdot\mathbf{E}) = (\mathbf{p}\cdot\mathbf{grad})\,\mathbf{E}^* \tag{2.16}$$

where **p** is removed from behind the symbol for the partial derivatives since it does not depend on the coordinates. If **E** does not depend on the coordinates either, *i.e.*, if **E** is constant in the region, **F** vanishes.

If a polarized dielectric is electrically equivalent to an assemblage of dipoles in a field, the corresponding energy may be found from Eq. (2.14). Writing $\mathbf{p}_v$ for the polarization per unit volume the energy associated with a volume dv is $-\mathbf{p}_v \cdot \mathbf{E}\,dv$ and integrating over the total volume occupied by the dielectric

$$U = -\int_v \mathbf{p}_v \cdot \mathbf{E}\,dv$$

As $\mathbf{E} = -\mathbf{grad}\,V$ and $\operatorname{div}(\mathbf{p}_v V) = \mathbf{p}_v \cdot \mathbf{grad}\,V + V\operatorname{div}\mathbf{p}_v$, the energy may be written

$$U = -\int_v V \operatorname{div}\mathbf{p}_v\,dv + \int_v \operatorname{div}(\mathbf{p}_v V)\,dv$$

Using the theorem of flux to transform the second integral into the integral of the normal component of $V\mathbf{p}_v$ over the bounding surface of the dielectric

$$U = -\int_v V \operatorname{div}\mathbf{p}_v\,dv + \int_s V\mathbf{p}_v \cdot \mathbf{ds} \tag{2.17}$$

This may be compared with the work that would have to be done to place the separated or *induced* charges constituting the dielectric in the positions that they occupy in the field. Writing V for the potential in the dielectric as above and q_v^i for the volume density of induced charge throughout the dielectric and q_s^i for the charge density induced on the surface

$$U = \int_v V q_v^i\,dv + \int_s V q_s^i\,ds \tag{2.18}$$

A multiplicative factor of $\frac{1}{2}$ does not appear in this expression as in Eq. (2.2) because the external potential is considered to be constant during the introduction of the dipoles. Comparing Eqs. (2.17) and (2.18), which are both expressions for the energy of the same configuration, it is evident that the volume and surface induced charge densities must be related to the polarization by the following equations:

$$q_v^i = -\operatorname{div}\mathbf{p}_v \tag{2.19}$$

$$\mathbf{p}_v \cdot \mathbf{ds} = q_s^i\,ds \qquad \text{or} \qquad p_n = q_s^i \tag{2.20}$$

* Equation (2.16) would be written in terms of the unit Cartesian vectors **i**, **j**, and **k** as

$$\mathbf{F} = \mathbf{i}\left(p_x\frac{\partial E_x}{\partial x} + p_y\frac{\partial E_x}{\partial y} + p_z\frac{\partial E_x}{\partial z}\right) + \mathbf{j}\left(p_x\frac{\partial E_y}{\partial x} + p_y\frac{\partial E_y}{\partial y} + p_z\frac{\partial E_y}{\partial z}\right) + \mathbf{k}\left(p_x\frac{\partial E_z}{\partial x} + p_y\frac{\partial E_z}{\partial y} + p_z\frac{\partial E_z}{\partial z}\right)$$

where p_n is the normal component of the polarization vector at the surface of the dielectric. If the polarization is uniform throughout the dielectric, $\mathbf{p}_v$ is independent of the coordinates and $q_v^i = 0$, *i.e.*, there is no net charge density within the dielectric. If, however, div $\mathbf{p}_v$ does not vanish the induced charge remaining in any volume after polarization is $-\int_v \text{div } \mathbf{p}_v \, dv$ or by the theorem of flux the surface integral of the inward normal component of $\mathbf{p}_v$ over the bounding surface of the volume. A surface charge appears even if the polarization is uniform and the surface density of this induced charge is equal to the normal component of the polarization at the surface.

FIG. 2.13.—Hypothetical closed surface for discussing the induced charge density in a dielectric. → represents the magnitude and direction of the polarization at a point.

The electric field is, of course, determined by all the charges present, those that are distributed throughout space or placed on conductors as well as the charge distribution induced by polarization. It is frequently a convenience to be able to discuss the electrical situation in terms of the charges that are placed in position without specific reference to those that are induced in any dielectric media that may be present. Writing q_v^t for the total charge density and using q_v for the charges placed throughout space or on conducting surfaces, $q_v^t = q_v + q_v^i$. In Eq. (1.28) q_v of course refers to total charge, hence in a dielectric medium

FIG. 2.14.—Charge induced on the surface of a dielectric due to polarization. →, polarization vector; -→, normal component of surface polarization.

$$\text{div } \mathbf{E} = \frac{q_v + q_v^i}{\kappa_0}$$

or eliminating q_v^i by means of Eq. (2.19)

$$\text{div } (\kappa_0 \mathbf{E} + \mathbf{p}_v) = q_v$$

The sum of the vectors $\kappa_0\mathbf{E}$ and $\mathbf{p}_v$ is called the *displacement* and written **D.** In general a medium is not isotropic and it is easier to distort the electrical lattice structure along one direction than along another. In this case $\mathbf{p}_v$ and $\mathbf{E}$ are not colinear. If, however, the medium is isotropic the vectors $\mathbf{p}_v$ and $\mathbf{E}$ and consequently their sum $\mathbf{D}$ are all in the same direction. In either case the differential equation remains

$$\text{div } \mathbf{D} = q_v \tag{2.21}$$

Little can be done with the general case but it is frequently justifiable to make certain simplifying assumptions. Limiting the discussion to

isotropic media and assuming that to a first approximation the molecular distortion is proportional to the electric field, *i.e.*, $\mathbf{p}_v = \kappa_0\chi\mathbf{E}$, where χ is a constant of the substance known as the *electric susceptibility*, $\mathbf{D}$ becomes

$$\mathbf{D} = \kappa_0\mathbf{E} + \mathbf{p}_v = \kappa_0(1 + \chi)\mathbf{E} \tag{2.22}$$

Equation (2.21) then becomes

$$\text{div } (1 + \chi)\mathbf{E} = \frac{q_v}{\kappa_0}$$

If in addition the medium is homogeneous so that χ does not depend on the coordinates of the point, $(1 + \chi)$ may be removed from the symbol representing the partial derivatives and it is evident on comparison with Eq. (1.28) that the electric field produced by the charge distribution q_v is less by the factor $1/(1 + \chi)$ than it would be if the dielectric were absent. It was seen, however, in Sec. 2.3 that the field between the plates of a charged and insulated condenser is decreased by the factor $1/\kappa$ on introducing a dielectric material; hence

FIG. 2.15.—(a) Displacement vector in an anisotropic medium. (b) Displacement vector in an isotropic medium.

$$\kappa = 1 + \chi$$

Equations (2.22) and (2.21) may then be written in terms of the dielectric constant as

$$\mathbf{D} = \kappa\kappa_0\mathbf{E} \tag{2.23}$$

and

$$\text{div } \kappa\kappa_0\mathbf{E} = q_v \tag{2.24}$$

Because of the linear relation assumed to hold between the polarization and the field, Eq. (2.24) is said to represent a *linear dielectric*. If in addition the dielectric is homogeneous, κ may be removed from the divergence symbol and the equation becomes

$$\text{div } \mathbf{E} = \frac{q_v}{\kappa\kappa_0} \tag{2.25}$$

This equation applies to gases, liquids, and isotropic solids so long as the applied electric field is not so great that the linear relation between the polarization and field ceases to hold.

If there is no volume distribution of placed charge, *i.e.*, if $q_v = 0$, it is evident that the differential equation for the field in a linear homo-

geneous isotropic dielectric is the same as that in free space. Thus Laplace's equation (1.30) holds for a charge-free dielectric. The electric charges giving rise to the field are excluded from the region under consideration by hypothetical closed surfaces that may coincide with dielectric or conductor boundaries. In this case the problem reduces to that of finding solutions of Laplace's equation which satisfy the boundary conditions at the interfaces between conductors and dielectrics or between two dielectrics. These conditions can be found from Eq. (2.21) or (2.24) and the relation between **E** and V [Eq. (1.13)] in an analogous manner to the derivation of Eq. (1.15). From Eq. (2.21) and the theorem of flux

$$\int_v \operatorname{div} \mathbf{D}\, dv = \int_s \mathbf{D} \cdot \mathbf{ds} = \int_v q_v\, dv = q \tag{2.26}$$

where q is the charge contained within the surface s. Applying this equation to a shallow pillbox surface containing an area ds of a con-

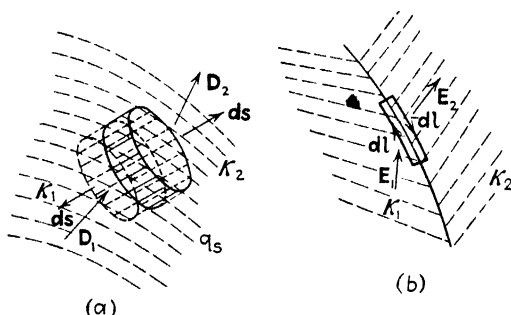



FIG. 2.16.—(a) Displacement vector at a dielectric boundary. (b) Electric field at a dielectric boundary.

ductor-dielectric boundary, the field vanishes in the conductor and the sides perpendicular to it make no contribution; hence the normal component of **D** times ds is equal to $q_s\, ds$, or

$$D_n = q_s \tag{2.27}$$

Therefore just outside the surface of the conductor, V is equal to the potential of the conductor and $E_n = -\dfrac{\partial V}{\partial n} = \dfrac{q_s}{\kappa\kappa_0}$, where n is the coordinate normal to the conducting surface.

At the interface between two dielectrics the same technique may be employed to determine the relation between the normal components of the displacement on the two sides. Consider a hypothetical pillbox surface enclosing the interface as shown at the left in Fig. 2.16. If the pillbox is made very shallow, the contribution to the integral over the curved sides may be made vanishingly small in comparison with that over the flat surfaces. If $\mathbf{D}_1$ is written for the displacement on the side

of the box in the medium of dielectric constant κ_1 and $\mathbf{D}_2$ for that in the medium of dielectric constant κ_2, Eq. (2.26) yields

$$\mathbf{D}_2 \cdot \mathbf{ds} - \mathbf{D}_1 \cdot \mathbf{ds} = q_s\, ds \qquad \text{or} \qquad D_{1n} - D_{2n} = q_s \tag{2.28}$$

The second condition is obtained from the fact that no work is done in taking a unit test charge around a closed path a portion of which lies in each medium. If the path is chosen so that the portions perpendicular to the interface are vanishingly short in comparison with those parallel to the surface as shown in Fig. 2.16

$$\mathbf{E}_1 \cdot \mathbf{dl} - \mathbf{E}_2 \cdot \mathbf{dl} = 0 \qquad \text{or} \qquad E_{1t} = E_{2t} \tag{2.29}$$

where the subscript t denotes the component of the field tangential to the boundary.[1] In the particular case in which no charge is placed on the interface between the dielectrics, $q_s = 0$ and the conditions are

$$\kappa_1 E_{1n} = \kappa_2 E_{2n} \qquad E_{1t} = E_{2t} \tag{2.30}$$

These equations represent a refraction of the lines of force at the interface, as shown in Fig. 2.17. If θ_1 is the angle between $\mathbf{E}$ and the normal to the surface in medium 1 and θ_2 is the similar angle in medium 2

$$\kappa_1 \mathbf{E}_1 \cos\theta_1 = \kappa_2 \mathbf{E}_2 \cos\theta_2$$

and

$$\mathbf{E}_1 \sin\theta_1 = \mathbf{E}_2 \sin\theta_2$$

Dividing one equation by the other the relation between the angles made by the lines of force in the two media with the normal to the surface is given by

$$\kappa_1 \cot\theta_1 = \kappa_2 \cot\theta_2 \tag{2.31}$$

$\kappa_1 \cot\theta_1 = \kappa_2 \cot\theta_2$

FIG. 2.17.—Refraction of a line of force at the boundary between two dielectric media.

2.5. Forces on Conductors and Dielectrics.—The work that must be done to establish a given charge configuration in the presence of dielectric media can be calculated as in Sec. 2.1. It is evident that the result is given by Eq. (2.2) where q_v represents the volume distribution of charge that is moved into position. However, Eq. (1.28) no longer applies, but it is replaced by Eq. (2.21). Thus

$$U = \tfrac{1}{2}\int V \operatorname{div} \mathbf{D}\, dv \tag{2.32}$$

Using the relation $\operatorname{div}(V\mathbf{D}) = V \operatorname{div}\mathbf{D} + \mathbf{D}\cdot\mathbf{grad}\, V$, the same argument may be used to change the integral to the surface integral of the normal component of $V\mathbf{D}$ and the volume integral of $\mathbf{D}\cdot\mathbf{grad}\, V$. If the charges

[1] In terms of the potential φ the boundary conditions are evidently $\varphi_1 = \varphi_2$ and $\kappa_1(\partial\varphi/\partial\mathbf{n})_1 - \kappa_2(\partial\varphi/\partial\mathbf{n})_2 = q_s$.

are localized, the surface integral vanishes on choosing a sufficiently distant surface and the volume integral may be written

$$U = \tfrac{1}{2}\int \mathbf{D} \cdot \mathbf{E}\, dv \tag{2.33}$$

by the aid of Eq. (1.13). Thus the energy per unit volume in terms of the field and displacement is given by $\frac{1}{2}\mathbf{D} \cdot \mathbf{E}$. If the medium is linear and isotropic, Eq. (2.33) reduces to

$$U = \frac{\kappa_0}{2}\int \kappa E^2\, dv \qquad \text{or} \qquad U = \frac{1}{2\kappa_0}\int \frac{D^2}{\kappa}\, dv \tag{2.34}$$

If in addition the medium is homogeneous, the dielectric constant may be removed from the integral sign.

These equations may be used to find the forces on conducting surfaces embedded in dielectric media or the forces exerted on the interface between two dielectrics. It will be assumed that the dielectrics are perfectly rigid and that the constraints permit any postulated displacement or that they are fluids that can alter their shape without having the mechanical forces which may thus be brought into play affect their dielectric constants. Actually the electric forces distort the dielectric giving rise to strains and changes in density. This phenomenon is known as *electrostriction*. It will here be neglected as its effects are too complicated for simple analysis.[1] Consider first an element ds of a conducting surface embedded in a dielectric which is given an outward displacement dl. The energy that disappears from the field is that which was contained in the volume $ds\, dl = dv$. Thus the decrease in energy is by Eq. (2.34) $\frac{1}{2}\kappa\kappa_0 E^2\, dv = \frac{1}{2}\,(D^2/\kappa\kappa_0)\, dv$. The force per unit area is this divided by ds and dl, or

$$F_s = \frac{1}{2}\kappa\kappa_0 E^2 = \frac{1}{2}\frac{D^2}{\kappa\kappa_0} = \frac{1}{2}\frac{q_s^2}{\kappa\kappa_0} \tag{2.35}$$

by Eq. (2.27). Thus the surface force is less by the factor $1/\kappa$ than it would be for the same surface charge density if the dielectric were absent. On the other hand, of course, if the potentials of the conductors are maintained the same when the dielectric is introduced, the charges q_s increase to a maximum of κq_s if the dielectric completely fills the region and the surface forces may thus increase by the factor κ.

On a dielectric in a field there are also body forces that can be obtained from Eq. (2.33). An important special situation is that in which the charges giving rise to $\mathbf{D}$ [Eq. (2.21)] remain fixed in position or the poten-

[1] See ABRAHAM AND BECKER, "Classical Electricity and Magnetism," Blackie & Son, Ltd., Glasgow, 1932; STRATTON, "Electromagnetic Theory," McGraw-Hill Book Company, Inc., New York, 1941; CADY, "Piezoelectricity," McGraw-Hill Book Company, Inc., New York, 1946.

tials of all the conductors which carry these charges are unaltered. In such cases Eq. (2.33) may be written in terms of the energy and field in the absence of the dielectric and the resulting polarization of the dielectric when in the field. Writing the subscript zero for the vectors in the presence of the charges but in the absence of any dielectric the increase in energy due to the introduction of the dielectric is

$$U - U_0 = \tfrac{1}{2}\int(\mathbf{D}\cdot\mathbf{E} - \mathbf{D}_0\cdot\mathbf{E}_0)\,dv$$

This can be simplified by writing it as

$$U = U_0 + \tfrac{1}{2}\int\{\mathbf{E}(\mathbf{D} - \mathbf{D}_0) + (\mathbf{E} - \mathbf{E}_0)\mathbf{D}_0\}\,dv$$

Now,

$$\begin{aligned}\int\mathbf{E}(\mathbf{D} - \mathbf{D}_0)\,dv &= -\int(\textbf{grad } V)\cdot(\mathbf{D} - \mathbf{D}_0)\,dv\\ &= -\int\text{div}\,[V(\mathbf{D} - \mathbf{D}_0)]\,dv + \int V\,\text{div}\,(\mathbf{D} - \mathbf{D}_0)\,dv\end{aligned}$$

The first term on the right vanishes by the argument used in deriving Eq. (2.33) and the second term vanishes as well if the charges are unchanged and hence: div $\mathbf{D}$ = div $\mathbf{D}_0 = q_v$. Therefore $U = U_0 + \tfrac{1}{2}\int\mathbf{D}_0(\mathbf{E} - \mathbf{E}_0)\,dv$. In a similar way it can be shown that $\tfrac{1}{2}\int\mathbf{E}_0(\mathbf{D} - \mathbf{D}_0)\,dv = 0$. Subtracting this last equation from that for U one obtains

$$U = U_0 + \tfrac{1}{2}\int(\mathbf{E}\cdot\mathbf{D}_0 - \mathbf{E}_0\cdot\mathbf{D})\,dv$$

For a linear isotropic dielectric, $\mathbf{D} = \kappa\kappa_0\mathbf{E}$ and:

$$\begin{aligned}U &= U_0 - \tfrac{1}{2}\kappa_0\int(\kappa - 1)\mathbf{E}\cdot\mathbf{E}_0\,dv\\ &= U_0 - \frac{1}{2}\int\frac{\kappa - 1}{\kappa}\mathbf{D}\cdot\mathbf{E}_0\,dv\\ &= U_0 - \tfrac{1}{2}\int\mathbf{p}_v\cdot\mathbf{E}_0\,dv \qquad (2.36)\end{aligned}$$

This form for the energy is a very convenient one for many calculations. From the fact that the angle between $\mathbf{p}_v$ and $\mathbf{E}_0$ is less than $\pi/2$, the integral is always positive and hence there is a decrease in electrostatic energy upon the introduction of a dielectric into a field. An instance of this has already been seen in Sec. 2.3. Equation (2.36) may also be used for calculating the energy change when a dielectric is introduced in the neighborhood of conductors that are maintained at constant potentials. This may be seen by an extension of the preceding analysis

in which the charges are permitted to change, in which case an additional term $\frac{1}{2}\int(q_s - q_{s0})(V - V_0)\,dv$ appears. This vanishes if either $q_s = q_{s0}$ or $V = V_0$. In the constant potential case, however, the batteries supply twice the negative change in electrostatic energy. Thus Eq. (2.36) for the electrostatic energy appears with a positive sign, but the equation for the total energy including that of the battery is, of course, unchanged.

2.6. Special Problems.—As an instance consider a U tube containing a liquid dielectric, one arm of which is placed between the plates of a condenser in such a way that the meniscus remains near the center of the region of uniform field. If the height of the meniscus increases from h_0 to h, the gravitational energy added to the system is $\rho gs(h - h_0)^2$, where ρ is the density of the fluid and s is the cross-sectional area of the tube. The equation for the electrostatic and gravitational potential energy is then

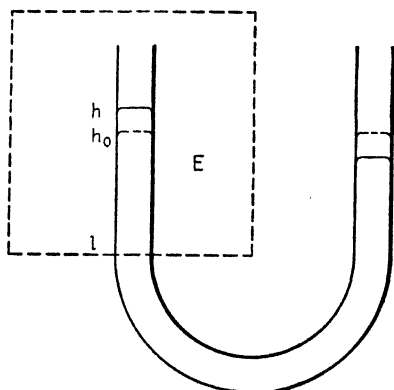



FIG. 2.18.—U-tube method of measuring the susceptibility of a liquid.

$$U = U_0 - \frac{1}{2}\int \mathbf{p}_v \cdot \mathbf{E}_0 + \rho gs(h - h_0)^2$$

The integral can be taken between an arbitrary constant lower limit, and the upper limit h and dv written as $s\,dx$, where x is the vertical coordinate. Equilibrium is determined by $dU/dh = 0$ or, since h is the upper limit of integration,

$$0 = -\frac{1}{2}s\mathbf{p}_v \cdot \mathbf{E}_0 + 2\rho gs(h - h_0)$$

or

$$h - h_0 = \frac{\mathbf{p}_v \cdot \mathbf{E}_0}{4\rho g}$$

In case a dielectric gas occupies the region above the meniscus, an extension of the above argument shows that

$$h - h_0 = \frac{(\mathbf{p}_{v2} - \mathbf{p}_{v1}) \cdot \mathbf{E}_0}{4\rho g} \tag{2.37}$$

where the subscripts 1 and 2 refer to gas and liquid respectively.

The height of rise depends on the geometry of the cross section. Three tractable cases are shown in Fig. 2.19. At a a cross section is shown that is long parallel to the field and narrow normal to it. The major areas that determine the vectors in the tube are the sides, and from

Eq. (2.30) $\mathbf{E}_1 = \mathbf{E}_2$ for the fields in the two media. If the dielectric is isotropic, $\mathbf{p}_v = (\kappa - 1)\kappa_0\mathbf{E}$ by Eq. (2.22) and from Eq. (2.37)

$$(h - h_0) = \frac{(\kappa_2 - \kappa_1)}{4\rho g}\kappa_0 E_0^2 \qquad \text{(case } a\text{)}$$

since $\mathbf{E}$ is the same as $\mathbf{E}_0$. In c the major dimension is normal to the field, and hence the predominant boundary condition from Eq. (2.30) is $\mathbf{D}_2 = \mathbf{D}_1$. For an isotropic dielectric $\mathbf{p}_v = (\kappa - 1)/\kappa\mathbf{D}$ and assuming the tube area to be so small as not to disturb the charge distribution on the

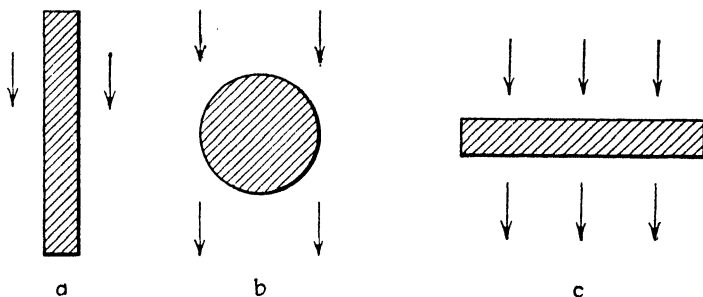



FIG. 2.19.—Cross sections of U-tube arm in Fig. 2.18.

plates at a great distance, the displacement is that previously existing in the gas, or $\mathbf{D} = \mathbf{D}_0 = \kappa_1\kappa_0\mathbf{E}_0$. Equation (2.37) yields[1]

$$(h - h_0) = \left(\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right)\frac{\kappa_1\kappa_0 E_0^2}{4\rho g} \qquad \text{(case } c\text{)}$$

In b the cross section is circular, and it is shown in a problem at the end of this chapter that the field within a small circular cylinder of dielectric constant is $2\mathbf{E}_0/(1 + \kappa)$ if $\mathbf{E}_0$ is the undisturbed field at a great distance. Thus in this case $\mathbf{p}_v = 2\kappa_0\mathbf{E}_0\dfrac{(\kappa - 1)}{(\kappa + 1)}$ and Eq. (2.37) yields

$$(h - h_0) = \frac{\kappa_2 - \kappa_1}{(\kappa_1 + 1)(\kappa_2 + 1)}\frac{\kappa_0 E_0^2}{\rho g} \qquad \text{(case } b\text{)}$$

Certain solid dielectric problems having simple geometries can be handled by a method similar to that of images. Consider a point charge q in a medium of dielectric constant κ_1 a distance a from an infinite-plane boundary with another medium of dielectric constant κ_2. The potential in medium 1 may be shown to be the same as that due to the charge q and an image charge equal to $[(\kappa_1 - \kappa_2)/(\kappa_1 + \kappa_2)]q$ placed a distance a on the other side of the boundary. The potential in medium 2 is

[1] If in this case the cross section is not small in comparison with the area between the plates but the long dimension extends the full width of the plates, the same result is seen to be obtained except κ_2 replaces κ_1 outside the parenthesis, since then $\mathbf{D} = \kappa_2\kappa_0\mathbf{E}_0$.

that which would be produced by a charge $[2\kappa_2/(\kappa_1 + \kappa_2)]q$ at the position actually occupied by q. The entire space is considered to be filled with the medium κ_2. This is illustrated in Fig. 2.20. Writing V_1 for the potential in medium 1 and V_2 for that in medium 2*

$$V_1 = \frac{q}{4\pi\kappa_0\kappa_1}\left\{[(x+a)^2 + y^2]^{-\frac{1}{2}} + \left(\frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2}\right)[(x-a)^2 + y^2]^{-\frac{1}{2}}\right\}$$

$$V_2 = \frac{q}{4\pi\kappa_0\kappa_2}\left(\frac{2\kappa_2}{\kappa_1 + \kappa_2}\right)[(x+a)^2 + y^2]^{-\frac{1}{2}}$$

These are solutions of Laplace's equation, as may be seen by substitution and they vanish for very large values of x and y. Also, V_1 reduces to

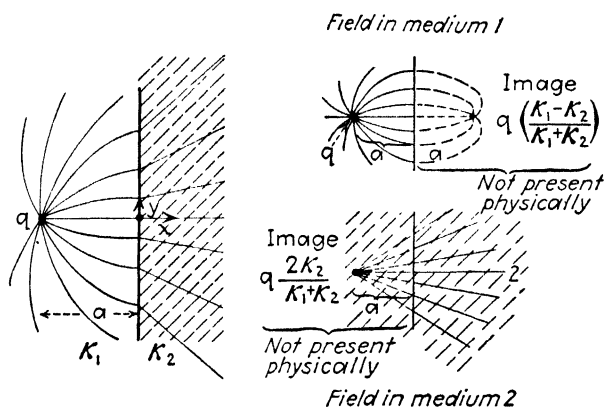



FIG. 2.20.—Method of images applied to a point charge in front of a plane dielectric boundary.

the potential of the single-point charge q in the neighborhood of the point $x = -a$, $y = 0$, where the charge is located. The fact that the necessary conditions are satisfied at the plane boundary may be verified by differentiation. In this way it may be seen that

$$\kappa_1 \frac{\partial V_1}{\partial x} = \kappa_2 \frac{\partial V_2}{\partial x} \qquad \text{and} \qquad \frac{\partial V_1}{\partial y} = \frac{\partial V_2}{\partial y}$$

at the boundary $(x = 0)$. The lines of force are indicated in Fig. 2.20 and the angles they make with the normal to the boundary are given by Eq. (2.31). The induced surface charge can be found from Eq. (2.28). The force on the charge q is that exerted by the charge $[(\kappa_1 - \kappa_2)/(\kappa_1 + \kappa_2)]q$ at the distance $2a$ or

$$F = \frac{1}{4\pi\kappa_0\kappa_1}\left(\frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2}\right)\frac{q^2}{4a^2}$$

* The actual derivation of the solutions of Laplace's equation used in this chapter will be found in the mathematical treatises previously referred to.

This is negative and q is attracted toward the boundary if $\kappa_1 < \kappa_2$, and it is positive indicating a repulsive force if $\kappa_1 > \kappa_2$.

Another readily calculable case is that of a sphere of radius a and dielectric constant κ_1 placed in a medium of dielectric constant κ_2 in which a uniform field E_0 previously existed. It may be shown in this case that the potentials in the two regions are given by

$$V_1 = -\left(\frac{3\kappa_2}{\kappa_1 + 2\kappa_2}\right)E_0 r \cos\theta$$

$$V_2 = -\left[1 - \frac{a^3}{r^3}\left(\frac{\kappa_1 - \kappa_2}{\kappa_1 + 2\kappa_2}\right)\right]E_0 r \cos\theta \tag{2.38}$$

in polar coordinates, where r is the distance from the center of the sphere and θ is the angle between the radius vector and the field. It may be shown by substitution that these satisfy Laplace's equation. At a great distance $(r >> a)$ $-\frac{\partial V_2}{\partial x} = E_0$, which is a necessary condition as the field must be unaltered at a great distance from the sphere. The conditions at the boundary between the dielectrics are also satisfied, for it is seen by differentiation that

$$\frac{1}{r}\frac{\partial V_1}{\partial \theta} = \frac{1}{r}\frac{\partial V_2}{\partial \theta}$$

and

$$\kappa_1 \frac{\partial V_1}{\partial r} = \kappa_2 \frac{\partial V_2}{\partial r}$$

FIG. 2.21.—Dielectric sphere in a uniform field.

when r is set equal to a. From the equations for the potentials it is evident that the field inside the sphere is uniform and equal to

$$\left(\frac{3\kappa_2}{\kappa_1 + 2\kappa_2}\right)E_0$$

If $\kappa_2 >> \kappa_1$, this approaches the value $\frac{3}{2}E_0$ and if $\kappa_1 >> \kappa_2$, it becomes small. The field outside is the original field plus that which would be produced by an electric dipole at the center of the sphere with a moment $p = 4\pi\kappa_2\kappa_0[(\kappa_1 - \kappa_2)/(\kappa_1 + 2\kappa_2)]a^3E_0$. The induced surface-charge density can be found from Eq. (2.28). By symmetry it is evident that there is no net body force on the sphere if the field is uniform.

2.7. Effective Molecular Field.—The electric field **E** is of course the force per unit charge on a test charge located at the point, but when dimensions of atomic magnitude are considered, **E** may no longer be considered uniform throughout the test charge.

Also, the average value of **E** over a region is not the effective value of the field under the influence of which a molecule is polarized, for the polarized molecule itself contributes to the average value of **E**. Nor is it possible to calculate this field on the basis of a real cavity in the dielectric. In the first place, the field in a cavity depends upon its shape. If it is in the form of a long narrow cylinder with its axis parallel to the field, the second of Eqs. (2.30), which states that the tangential components of **E** are the same in the dielectric and in the cavity, implies that near the center of the cavity the field is the same as in the dielectric. If the cavity is in the form of a shallow disk with its axis parallel to the field, the fact that the normal component of **D** is continuous implies that the field within the cavity and near its center is κ times the field in the dielectric, where κ is the dielectric constant. If the medium is completely isotropic, *i.e.*, if there is no permanent polarization associated with the constituent molecules and the molecules themselves may be considered as minute spheres, it is reasonable to calculate the effect of the polarization of neighboring molecules on one another on the assumption of spherical symmetry. However, if a spherical cavity were actually made in the dielectric the lines of force would be refracted at the boundary and the field would be given by Eq. (2.32). This, however, is not the situation that exists when the molecule is present, for then the lines of force continue straight through the region occupied by the molecule.

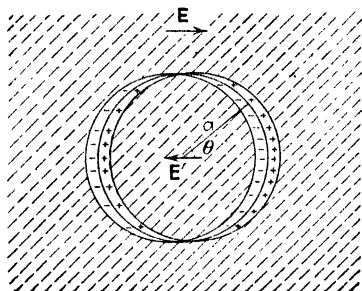



FIG. 2.22.—Determination of the effective atomic electric field in a nonpolar medium.

The simplest method of achieving the desired result is to consider a hypothetical spherical surface in the dielectric within which the molecule is located. Within this sphere are the charges associated with the molecule or a spherically symmetrical group of molecules. If this region is assumed to be removed without making any change in the external field, the charges remaining on the inner walls of the cavity so left will produce a field in the opposite direction to **E**. This field, say $-\mathbf{E}'$, may be calculated by the aid of Eq. (2.20). Since the surface density of charge is proportional to the normal component of the polarization, q_s equals $p_v \cos\theta$, where the angle θ is indicated in Fig. 2.22. If the cavity is of radius a, the field at the center, $-\mathbf{E}'$, is given by

$$\begin{aligned} -\mathbf{E}' &= \frac{-\mathbf{p}_v}{4\pi\kappa_0 a^2}\int_0^\pi \cos\theta\cos\theta\, 2\pi a^2 \sin\theta\, d\theta \\ &= \frac{+\mathbf{p}_v}{2\kappa_0}\int_0^\pi \cos^2\theta d(\cos\theta) \\ &= \frac{-\mathbf{p}_v}{3\kappa_0} \end{aligned} \tag{2.39}$$

This is the effective depolarizing field produced by a spherically symmetrical molecule. Since the total field including this is **E**, the effective field which acts to polarize the molecule is $\mathbf{E} - \mathbf{E}'$, or

$$\mathbf{E}_{\text{eff.}} = \mathbf{E} + \frac{\mathbf{p}_v}{3\kappa_0} \tag{2.40}$$

Thus the field effective on the molecule is greater than **E** by one-third of the polarization per unit volume divided by κ_0.

Equation (2.40) for the local polarizing field is frequently known as the Clausius-Mosotti formula. If a linear relation exists between the field and the polarization, $\mathbf{p}_v$ may be written in terms of $\mathbf{E}$ and the susceptibility or dielectric constant as

$$\mathbf{p}_v = \chi\kappa_0\mathbf{E} = \kappa_0(\kappa - 1)\mathbf{E}$$

Eliminating $\mathbf{E}$, Eq. (2.38) becomes

$$\mathbf{p}_v = 3\kappa_0\left(\frac{\kappa - 1}{\kappa + 2}\right)\mathbf{E}_{\text{eff.}}$$

The coefficient of $\mathbf{E}_{\text{eff.}}$ is of particular interest for atomic purposes, for this yields the actual molecular distortion per unit effective field. If α is written for the molecular electric moment induced per unit field $\mathbf{p}_v = \alpha n\mathbf{E}_{\text{eff.}}$, where n is the number of molecules per unit volume. In terms of the molecular weight, Avogadro's number and the density $n = N\rho/M$ and $\mathbf{p}_v = \alpha N\rho\mathbf{E}_{\text{eff.}}/M$. Inserting this value of $\mathbf{p}_v$ in the above equation

$$\alpha = \frac{3M\kappa_0}{N\rho}\left(\frac{\kappa - 1}{\kappa + 2}\right) \tag{2.41}$$

This equation determines the molecular polarizability in terms of the other measurable physical constants.[1] Being a molecular constant, α should, of course, be independent of the density or state of aggregation and within the limits imposed in this discussion it is found to be so. It will be seen in the discussion of radiation that the index of refraction of a medium for electromagnetic radiation is related to the dielectric constant and Eq. (2.41) has been tested also by measurements on the index of refraction. In the case of a gas, κ is very nearly unity and the equation for $\mathbf{p}_v$ becomes $\mathbf{p}_v = \kappa_0\chi\mathbf{E}_{\text{eff.}}$. Thus the effective electric field polarizing a molecule is the same to this approximation as $\mathbf{E}$ itself, the contribution of neighboring molecules being negligible. The information that is obtained by measuring dielectric constants and applying Eq. (2.41) to calculate the molecular constant α in those cases where the assumptions on which this discussion rests are valid is of great importance in the field of atomic physics. For a more complete discussion of this subject the reader is referred to special treatises in the field.[2]

Problems

1. A flat metal plate 0.01 m.^{2} in area is suspended from a balance arm so that it is parallel to a horizontal metal plate and 2 mm. above it. Show that a mass of 0.0113 gm. must be added to the other arm of the balance if the plate is to remain in position when a difference of potential of 100 volts is applied between the plates. If the space between the plates is filled with oil having a dielectric constant of 2, what mass would have had to be added to the other arm?

2. A condenser of capacity 10^{-10} farad is connected across the terminals of a

[1] In the system of units here adopted N would be the number of molecules in a mass of M kg. (6.06×10^{26}) and ρ is the density in kilograms per cubic meter. Alternatively κ_0 may be changed to centimeters (8.85×10^{-14} farad per centimeter) in which case M, N, and ρ have their cgs. values and α is in coulomb-centimeters per volt per centimeter.

[2] VAN VLECK, "Electric and Magnetic Susceptibilities," Oxford University Press, New York, 1932; DEBYE, "Polar Molecules," Reinhold Publishing Corporation, New York, 1929; SMYTH, "Dielectric Constant and Molecular Structure," Reinhold Publishing Corporation, New York, 1931.

charged electrostatic voltmeter and the deflection of the instrument is found to decrease 10 per cent. Show that the capacity of the meter is 9×10^{-10} farad.

3. A bubble of radius a is formed from a soap solution having a surface tension T. The external atmospheric pressure is P. (The internal pressure is P plus the contribution made by the surface tension of the film which may be shown to be $4T/a$.) If the bubble is raised to a potential V by being touched with a wire from a static machine, show that the radius of the bubble increases to r, where r is given by

$$P(r^3 - a^3) + 4T(r^2 - a^2) - \frac{\kappa_0 V^2}{2} r = 0$$

4. In the discussion of the quadrant electrometer in the text it was assumed that the coefficients of capacity and induction were given to a sufficient approximation by $a_{ij} + b_{ij}\theta$. To the next approximation they would be given by $a_{ij} + b_{ij}\theta + d_{ij}\theta^2$. Assuming this approximation and the symmetry postulated in the text, derive the expression for the angular deflection in the heterostatic case. Compare the sensitivity $(d\theta/dV)$ with that given by the first approximation.

5. With what force per unit area do the plates of a parallel-plate condenser attract one another if their separation is 1 mm. and the potential difference between them is 100 volts assuming that the region between them has a dielectric constant of unity?

6. If the condenser of the preceding problem is charged to the potential difference of 100 volts and then disconnected from the battery and the region between the plates is filled with paraffin oil having a dielectric constant of 2.3, what is the force of attraction per unit area of the plates? If the plates are then reconnected to the 100-volt battery, what is the force per unit area between them?

7. If the condenser plates of the preceding problem have an area of 0.01 m.^{2}, what is the increase in energy of the condenser at reconnection?

8. The metal bottom and piston of an airtight cylinder with insulating walls form the plates of a parallel-plate condenser. If the separation and pressure are initially d and p, respectively, when the condenser is uncharged, show that the fractional decrease in separation, f, when a potential difference V is applied, is given by

$$f(1 - f) = \frac{\kappa_0}{2p}\left(\frac{V}{d}\right)^2$$

9. Two condensers of capacity C_1 and C_2 possessing initially charges q_1 and q_2, respectively, are connected in parallel. Show that there is a loss of electrostatic energy amounting to

$$\frac{(C_2 q_1 - C_1 q_2)^2}{2C_1 C_2 (C_1 + C_2)}$$

In what form does this energy appear?

10. Two condensers, 1 and 2, have capacities C_1 and C_2, respectively. Condenser 1 is charged by a battery and after the battery is removed, a spark is drawn between the plates. 1 is then charged as before and 1 and 2 are connected in parallel, a spark appearing as the connection is made. 1 and 2 are then separated and each discharged by a spark. Show that the energies of the four sparks are in the ratio

$$(C_1 + C_2)^2 : C_2(C_1 + C_2) : C_1^2 : C_1 C_2$$

11. An electrostatic voltmeter is constructed in the form of a rotary variable condenser with n movable plates. The electrostatic torque is opposed by the restoring torque of a fiber which is given by $k\theta$, where θ is the angle of entry of the plates. If d

is the distance from a fixed to a movable plate and V is the potential difference between the two sets, show that the electrostatic torque is given by $(n\kappa_0/2d)r^2V^2$, where r is the radius vector to the rim of the movable plates at the point of entry. If the equation of the rim of the plates is $r^2\theta$ = const., find the equation for the angular deflection in terms of the potential difference.

12. A small hemispherical boss of radius a is raised on the inside of one plate of a parallel-plate condenser. If the potential difference between the plates is V and their separation is d $(d >> a)$, show that the electrical force tending to pull the boss from the plate is $\frac{3}{4}\pi\kappa_0(a^2/d^2)V^2$. (The potential is similar to that of Prob. 25, Chap. I.)

13. A line charge of linear density q_l is placed in a medium of dielectric constant κ_1 parallel to, and at a distance a from, the plane boundary with another medium of dielectric constant κ_2. Find the potentials in the two media and show that the force per unit length on the line charge is given by

$$F_l = \frac{q_l^2}{4\pi\kappa_0\kappa_1}\left(\frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2}\right)\frac{1}{a}$$

14. An infinite circular cylinder of dielectric constant κ_1 is placed in a medium of dielectric constant κ_2 with its axis perpendicular to a uniform field E_0 previously existing in the medium. Show that the potentials within and without the cylinder are given respectively by

$$V_1 = -\frac{2\kappa_2}{\kappa_1 + \kappa_2}E_0\, r\cos\theta$$

$$V_2 = -\left(1 - \frac{a^2}{r^2}\frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2}\right)E_0\, r\cos\theta$$

where r is the axial vector from the center of the cylinder and θ is the angle it makes with the field.

15. A small sphere of susceptibility χ and radius a is placed at a great distance r from a conducting sphere of radius b which is maintained at a potential V. Show that the force with which the dielectric sphere is attracted to the conducting one is

$$F = \frac{8\pi\kappa_0 a^3b^2V^2\chi}{(\chi + 3)r^3}$$

if the susceptibility of the intervening medium is negligible and $a << r$.

16. Show that if the sphere of the preceding problem has a density ρ and is suspended by a light fiber in the earth's gravitational field in the horizontal plane of the center of the conducting sphere, its susceptibility is given in terms of the angular displacement, θ, of the fiber from the vertical by

$$\chi = \frac{2\rho g r^5\tan\theta}{6\kappa_0 b^2V^2 - \rho g r^5\tan\theta}$$

17. A condenser is formed of two concentric spherical conducting shells of radii a and b. If the medium between the spheres has a dielectric constant κ_1 from a to r and κ_2 from r to b, show that the capacity of the condenser is given by

$$C = 4\pi\kappa_0\left[\frac{1}{\kappa_1 a} - \frac{1}{\kappa_2 b} + \frac{1}{r}\left(\frac{1}{\kappa_2} - \frac{1}{\kappa_1}\right)\right]^{-1}$$

18. A condenser is formed of two concentric conducting spherical shells of radii a and b. The inner sphere of radius a receives a thin coat of a material with a dielectric constant κ. Show that the capacity is changed approximately by

$$4\pi\kappa_0\left[\frac{b^2(\kappa - 1)}{\kappa(b - a)^2}\right]t$$

where t is the thickness of the dielectric coat.

19. The dielectric constant of the material between the plates of a parallel-plate condenser varies uniformly from one plate to the other. If κ_1 and κ_2 are its values at the two plates and d is the plate separation, show that the capacity per unit area is

$$C_s = \frac{\kappa_0}{d}\frac{\kappa_2 - \kappa_1}{\log_e (\kappa_2/\kappa_1)}$$

20. Two square conducting plates of length a on a side are placed parallel to one another a distance d apart $(d << a)$. A slab of material of dielectric constant κ, which is also a on a side but of thickness t, is inserted parallel to the edges of the plates. Assuming the dielectric constant of air to be unity and neglecting the effects at the edges of the plates show that the force with which the slab is drawn between the plates is given by

$$\frac{\kappa_0}{2}\frac{(\kappa - 1)t}{(d - t)\kappa + t}\frac{a}{d}V^2$$

where V is the potential difference between the plates.

21. Two coaxial cylindrical surfaces of radii a and b are lowered vertically into a liquid dielectric. If the liquid rises a distance h between the plates when a potential V is established between them, show that the susceptibility of the liquid is given by

$$\chi = \frac{(b^2 - a^2)\log_e (b/a)h\rho g}{\kappa_0 V^2}$$

where ρ is the density of the liquid and the susceptibility of air is neglected.

22. Find the times required for the plates of a parallel-plate condenser, of mass m_s per unit area, to come into contact when released from a separation x_0 when (a) the plates are connected momentarily to a source of potential V_0 and the stops removed, (b) they are connected permanently to a source of potential V_0 and the stops removed.

23. A point charge q is placed at the center of a spherical hole of radius a in a block of material of dielectric constant κ. Find the potential at all points. Prove that the sum of the induced charges and the original charge is q/κ, independent of a.

24. Two vertical conducting plates form a wedge of small angle α with one another and are maintained at a potential difference V. If the bottom edges of the plates are placed in a bath of oil of density ρ and dielectric constant κ, show that the oil rises to a height h at a distance x from the vertical intersection of the plates, where

$$h = \frac{1}{2}\frac{(\kappa - 1)\kappa_0 V^2}{(\alpha x)^2\rho g}$$

25. A line charge is placed at a distance d on one side of the plane interface separating two dielectric media. If it is in the medium of dielectric constant κ_1, the force per unit length acting upon it is found to be F_1; and if it is in the medium of dielectric constant κ_2, the force is found to be F_2. Show that

$$F_1\kappa_1 = F_2\kappa_2$$

26. A long block of dielectric material of dielectric constant κ is of such a thickness that it will just slide between two long vertical condenser plates of separation d and height L. The block rests on a table and the condenser plates are lowered until they are just above the block. The block is then held down on the table, and the condenser is charged by momentarily connecting it to a battery of potential V_0. The battery connections are then removed and the block released. Show that it will rise if $\dfrac{(\kappa - 1)V_0^2}{2m_l\, dg} > 1$, where m_l is the mass of the block per unit length parallel to the plates and g is the acceleration of gravity. Show that if this condition is fulfilled, the block will rise to a maximum height h given by

$$h = \frac{L}{\kappa - 1}\left(\frac{(\kappa - 1)\kappa_0 V_0^2}{2m_l\, dg} - 1\right)$$

Describe the subsequent motion of the block. The dimensions are such that edge effects are assumed to be negligible.

27. Two identical dipoles are mounted so that one lies in the perpendicular bisector of the other. Show that the torque tending to rotate one dipole about its axis is twice that tending to rotate the other about its axis.

28. A conducting sphere of radius a rests snugly in a depression in the apex of a vertical cone of insulating material of half angle α and dielectric constant κ. If the conical walls were projected the apex of the cone would be at the center of the sphere. Assuming this system is far removed from all other objects and a charge q is placed upon the sphere, show that it is attracted toward the cone by a force which is equal to

$$\frac{(\kappa - 1)\sin^2\alpha\, q^2}{32\pi\kappa_0 a^2}$$

29. A pair of horizontal coaxial cylindrical conductors are formed by a wire of radius a within a tube of inner radius b. The wire is supported by a long overhead wedge of insulating material of dielectric constant κ which fits the wire and tube snugly. The sides of the wedge are planes through the axis at equal inclinations α with the vertical. If a potential difference V is maintained between wire and tube, show that the wire will stay in place without being fastened to the wedge if its density ρ is given by

$$\rho = \frac{V^2(\kappa - 1)\kappa_0 \sin\alpha}{g\pi a^3 \log_e(2b/a)}$$

30. A conducting sphere of density ρ' floats in a liquid of density ρ ($\rho > 2\rho'$) and dielectric constant κ. The sphere is given a charge q such that it sinks until just half of its area is submerged. Show that q is given by

$$q^2 = \frac{(4\pi)^2\kappa_0(\kappa + 1)^2 a^5 g(\rho - 2\rho')}{3(\kappa - 1)}$$

CHAPTER III

PHYSICAL CHARACTERISTICS OF DIELECTRICS AND CONDUCTORS

3.1. Gaseous Dielectrics.—The discussion of the preceding chapters has been largely concerned with the limiting ideal cases of materials that were either perfect conductors in which no electric field could exist or perfect insulators in which there was no net motion of charge. While these ideals are closely approached by certain physical substances, they can never be actually realized. Many of the most important physical phenomena are associated with the divergence of real substances from their ideal prototypes and the dependence of their electrical characteristics on physical parameters such as temperature, pressure, strain, etc. In the gaseous state the molecules are separated by relatively great distances from one another and their mutual interaction may be neglected to a first approximation. For this reason a gas is the simplest type of substance to discuss quantitatively from a molecular point of view. It is homogeneous and isotropic and if the electric field is not too great, the electrical distortion that it produces in a molecule is proportional to the field. Thus such a medium is practically an ideal one in the sense of the preceding chapter, and in view of the relatively great separation of the molecules at ordinary pressures the effective polarizing field is practically the same as the applied field **E**. The molecular polarization constant α is then given by

$$n\alpha = \frac{\mathbf{P}_v}{\mathbf{E}} = \kappa_0\chi = \kappa_0(\kappa - 1) \tag{3.1}$$

α may be determined from this equation if the susceptibility or dielectric constant and the number of molecules per unit volume, n, are known.

Since n is proportional to the density, it is directly proportional to the pressure and inversely proportional to the temperature if the perfect-gas law is obeyed. The susceptibility is found experimentally to be proportional to the pressure over a wide range and for certain types of gases it is also found to be inversely proportional to the temperature. For other types of gases, however, χ varies more rapidly with the temperature and it is evident that the assumptions on which Eq. (3.1) is based are inadequate for them. Other molecular evidence suggests that the molecules of gases of this type are not really spherically symmetrical

but the electrical centers of positive and negative charge are normally displaced from one another, resulting in a permanent electrical dipole associated with each molecule. In consequence the effective molecular electric moment $\alpha\mathbf{E}$ must be supplemented by the mean component of the permanent electric dipole in the direction of the field. It may be shown that this leads on simple assumptions to an effective polarization

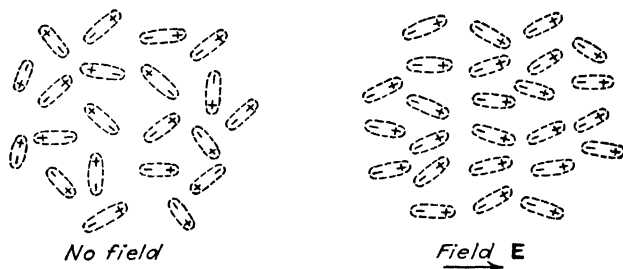



FIG. 3.1.—Schematic representation of the instantaneous orientation of the molecules of a polar gas under the influence of an electric field.

constant, $\alpha' = \left(\alpha + \frac{\beta}{T}\right)$, where T is the absolute temperature and α and β are molecular constants independent of T.* α is a measure of the

* It may be shown from a statistical analysis of a gaseous assemblage of molecules in thermal equilibrium under the influence of an external force field, that the probability of any one molecule having an energy U due to its position or orientation in the field is proportional to $e^{-\frac{U}{kT}}$, where k is Boltzmann's constant and T is the absolute temperature. The energy of a polar molecule of moment $\mathbf{p}$ in an electric field $\mathbf{E}$ is $-\mathbf{p}\cdot\mathbf{E} = U$. Also, the component of $\mathbf{p}$ in the direction of the field of a molecule, whose moment lies in the conical solid angle $d\omega$ of Fig. 3.2, is $p\cos\theta$. The net contribution to the polarization perpendicular to the field will vanish from symmetry. The average contribution of the orientation of molecules to the polarization in the direction of the field is the integral of the probability of a molecule making an angle between θ and $\theta + d\theta$ with the field, times its contribution to the polarization in that direction when it does so, over all values of θ. If the number of molecules with their moments lying in the solid angle $d\omega = 2\pi\sin\theta\, d\theta$ is say $Ne^{\left(\frac{pE}{kT}\right)\cos\theta}$ the mean effective polarization in the direction of the field is given by

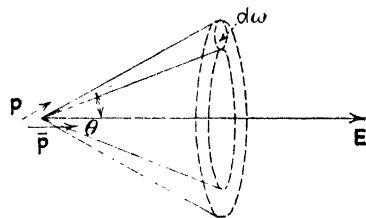



FIG. 3.2.—Calculation of the mean effective polarization induced by the orientation of polar molecules in a field.

$$\bar{p} = \frac{\int Ne^{x\cos\theta} p\cos\theta\, d\omega}{\int Ne^{x\cos\theta}\, d\omega}$$

where x is written for pE/kT. Writing $d\omega$ as $-2\pi d(\cos\theta)$, the denominator integrates to $N(e^x - e^{-x})/x$ since the limits of integration for $\cos\theta$ are $+1$ and -1. The

molecular distortion produced by the field and β involves the magnitude of the permanent electric dipole associated with a molecule. These permanent dipoles are normally oriented at random, making no net contribution to the polarization, but they tend to line up in the presence of a field and then do make a net contribution to the volume polarization. Molecules having a permanent electric moment are known as *polar molecules* and a measurement of α' as a function of the temperature may be used to calculate both the polarizability α and the permanent electric moment involved in β. In practice α' is measured at a series of temperatures and plotted as a function of $1/T$. The extrapolated intercept with the axis of infinite T yields α and the slope of the line is equal to β. α' is, of course, a constant equal to α and the slope β is zero for a nonpolar gas. Typical polar gases are HCl, HBr, SO_2, H_2O, NH_3, etc. Molecules such as H_2, N_2, O_2, and those of the rare gases as well as symmetrical organic ones such as CH_4, CCl_4, etc., have no permanent electric moment associated with them.

The dielectric constants of a few representative gases at 0°C. and normal atmospheric pressure are given in Table I. These values are

TABLE I

Gas	Dielectric constant	Susceptibility
Hydrogen	1.000264	2.64×10^{-4}
Air	1.000590	5.90×10^{-4}
Methane	1.000944	9.44×10^{-4}
Carbon dioxide	1.000985	9.85×10^{-4}

obtained by making precision capacity measurements and it is evident from the smallness of the susceptibility that the effect of the presence

numerator is seen to be p times the partial derivative of the denominator with respect to x, hence

$$\frac{\bar{p}}{p} = \frac{e^x + e^{-x}}{e^x - e^{-x}} - \frac{1}{x} = \coth x - \frac{1}{x}$$

This is the Langevin-Debye expression for the ratio of the mean effective component of the electric moment of a molecule to its actual dipole moment for a polar gas in a field E at a temperature T. For values of the field and temperature ordinarily encountered x is small. In this case coth x reduces to $\frac{1}{x} + \frac{x}{3}$ to this approximation and $\bar{p} = Ep^2/3kT$. But this is the effective contribution of the permanent moment of a polar molecule induced by the field E which is $E\beta/T$ of the previous discussion, or

$$\beta = \frac{p^2}{3k}$$

Thus an experimental determination of β yields the value of the permanent electric moment associated with the molecule.

of a gas between condenser plates may be neglected without introducing an error greater than 0.1 per cent. A calculation of α from Eq. (3.1) also shows that the molecular distortion produced by ordinary electric fields is extremely small. The values of the permanent electric moments associated with polar molecules which are determined by measuring the temperature variation of the dielectric constant of these gases are also very small in terms of ordinary units. However, if appropriate atomic units are chosen, such as 10^{-8} cm. as the unit of length and the electron as the unit of charge, these permanent moments are of the order of unity, as might be expected.

Gases approach most nearly to the ideal of perfect dielectrics. Though a few free electronic charges are produced per cubic centimeter per second in air owing to the presence of radioactive substances, cosmic rays, etc., the currents to which they give rise are extremely small. However, if the electric fields become large, this residual ionization is enhanced by various processes that will be considered further in Chap. VIII. At a field strength of the order of 3×10^6 volts per meter corona discharge sets in and the gas loses its insulating properties. The critical field for which such a cataclysmic effect takes place and for which the dielectric loses its insulating properties is known as the *dielectric strength* of the substance. The dielectric strength of a gas is approximately proportional to the pressure, and as the dielectric constant varies in the same manner, it is advantageous to increase the pressure of a gaseous dielectric. In the cases of solid and liquid dielectrics the orientation and distortion of the molecules and molecular aggregates that are produced by the application of a field result in a generation of heat through internal frictional effects. This type of loss of electrical energy is known as *dielectric loss* and one of the chief advantages of a gaseous dielectric is that these losses are in general negligible. This is of particular importance in the case of rapidly alternating electric fields as the power loss increases with the frequency of alternation.[1]

3.2. Liquid and Solid Dielectrics.—*Liquids.*—In the liquid state the mean distance apart of the molecules of a substance is of the order of magnitude of the dimensions of the molecules themselves. The mutual interaction of the molecules is therefore of great importance in determining the electrical properties of a substance. The Clausius-Mosotti relation can be used to calculate the dependence of the dielectric constant on density for nonpolar liquids and even for solutions of polar molecules in nonpolar liquids, but the fundamental assumption of microscopic isotropy is violated in the case of polar liquids. The dielectric constants

[1] References: Dielectric theory, VANVLECK, SMYTH, *loc. cit.;* Breakdown, PEEK: "Dielectric Phenomena in High Voltage Engineering," McGraw-Hill Book Company, Inc., New York, 1929.

of pure liquids vary from about 2 for the petroleum oils to about 80 for water. Certain colloidal suspensions of particles with large dipole moments display extremely high dielectric constants, of the order of 10^2 or 10^3.

Liquids are not as good insulators as gases or certain types of solids. A certain small proportion of the molecules even of a pure liquid are not electrically neutral but are positive or negative ions. These move through the liquid under the influence of a field and result in a net transfer of charge from one region to another. The resistivities listed in Table II which are proportional to the resistance offered by the liquid to the transfer of charge are large in comparison with good conductors but small compared to gases or certain solid dielectrics. Also the existence of any solid material in suspension results in a decreased resistance and in general a lowered dielectric strength. Colloidal particles in general possess a net charge and their motion through the liquid under

TABLE II

Liquid	Dielectric constant	Resistivity, ohm-meters
Petroleum oil	2.2	10^{11} (100°C.)
Linseed oil	3.3	6×10^9 (100°C.)
Castor oil	4.3	6×10^8 (100°C.)
Ethyl alcohol	25.8	3×10^3 (18°C.)
Methyl alcohol	31.2	1.4×10^3 (18°C.)
Water	81 07	5×10^3 (18°C.)

the influence of a field is known as *cataphoresis*. It will be observed from Table II that large dielectric constants are associated in general with low resistivities. Dielectric losses are also in general large for liquids with large dielectric constants. The paraffin oils are most widely used as commercial liquid dielectrics. Though their dielectric constant is low, they are chemically stable and relatively inexpensive to obtain in suitable purity. The dielectric strength is high for the pure oil, though it is greatly reduced by the presence of moisture. The dielectric strength of an oil is generally given in terms of the number of kilovolts which will cause breakdown between 1-in. disks 0.1 in. apart submerged in the oil. The strength of commercial insulating oils lies in the range from 30 to 50 in these units. A liquid dielectric reforms after a breakdown, which is an advantage not possessed by a solid dielectric. However, the chemical decomposition that frequently takes place impairs its insulating properties. The electrical adjustment of the liquid to the strain produced by the field is not instantaneous as in the case of gases but requires in general a small fraction of a second. Breakdown is

found to depend on the duration of the applied potential. Also the dielectric strength in general decreases with increasing temperature. Dielectric losses in a liquid are not negligible and the heating thus produced decreases the dielectric strength. Dielectric losses are generally given in terms of the *power factor* which represents approximately the fraction of the energy of the charged condenser that is lost in producing the polarization of the dielectric. For mineral oil this factor is about 4×10^{-4}, while for castor oil, for instance, it is about 20 times as great.[1]

Isotropic Solids.—Solid dielectrics are used for insulation and mechanical support in almost every type of electrical device. A wide variety of substances find application in different types of work. Mechanical properties are frequently of greater importance than electrical ones in ordinary service. Where mechanical rigidity is desired, wood, glass, porcelain, and synthetic fibers and resins are widely used, while protective coverings and flexible supports are provided by enamels, varnishes, impregnated fabrics, and rubber. Where very high insulation is necessary, as in electrostatic work, sulphur, amber, polyethylene, and quartz are used. In the temperature range below about 1000°C. mineral fibers and ceramic materials are most frequently encountered, and quartz can be used nearly up to its melting point at 1700°C., though the resistivity of all these substances decreases with increasing temperature. Surface leakage of electricity is of importance for many purposes and substances differ widely in this characteristic. Also the effect of humidity is different for different substances. For instance, dry quartz is an excellent insulator, but its surface resistivity decreases rapidly with increasing humidity whereas the resistivity of paraffin waxes is but little affected. Materials with high dielectric constants and of high dielectric strength such as impregnated paper, glass, and mica are used as separating media for condensers.

The majority of these solid materials have a definite crystal structure, which means that they are not isotropic on an atomic scale. However, the larger blocks of these materials are generally composed of randomly oriented microcrystalline aggregates so that on a macroscopic scale they are sensibly isotropic. Hence it is significant to attribute to them a scalar dielectric constant. However, the dielectric constants for classes of substances such as porcelains, glasses, micas, etc., are not well defined but vary between wide limits. The values given in Table III are to be regarded merely as representative. Also these substances are seldom homogeneous and as will be seen later this implies that surface and volume charges will appear when they are placed in an electric field.

[1] For a more complete discussion see Gemant, "Liquid Dielectrics," John Wiley & Sons, Inc., New York, 1933.

The internal molecular adjustments require a longer time in solid media than in liquids and there are generally appreciable time lags before a state of electrical equilibrium is reached. Bound charges and relaxation times of many minutes greatly complicate the phenomena that are observed in condensers having solid dielectrics. Breakdown also depends on the way in which the field is applied and upon its duration. Internal frictional effects are of much greater importance for solids than for liquids and in general the dielectric losses are considerably higher. All of these phenomena depend on external physical circumstances such as tempera-

TABLE III[1]

Substance	Resistivity, ohm-meters			Dielectric constant		Power factor		Dielectric strength, kv./mm.[6]
	Volume	Surface, 60 % R.H.[2]	Surface, 90 % R.H.	L.F.[3]	H.F.[4]	M.F.[5]	H.F.	
Amber	5×10^{14}	...	3×10^{14}	2.8				
Bakelite	10^{9}	10^{12}	10^{10}	4.9	3.7	0.03	0.04	24
Beeswax	...	...	...	2.7	2.3	0.02	0.005	
Cellulose acetate	...	...	...	3.8	3.2	0.01	0.03	10
Ceresin	$>5 \times 10^{16}$	...	$>8 \times 10^{16}$	2.2				
Mica	5×10^{11}	10^{10}	10^{8}	5.4	5.4	0.002	<0.0003	10 to 100
Micalex	...	...	...	7.1	6.9	0.006	0.004	38
Neoprene	10^{11}	10^{15}	10^{9}	6.9	4.1	0.01	0.04	12
Paper	10^{10}	10^{11}	10^{9}	3.7	....	0.009	...	16
Parowax	10^{14}	...	6×10^{13}	1.9	....	...	...	11 5
Plexiglass	10^{14}	10^{16}	10^{16}	3.4	2.6	0.06	0.006	40
Polyethylene	$>5 \times 10^{14}$	$>5 \times 10^{16}$	$>3 \times 10^{10}$	2.26	2.26	<0.0006	0.0004	50
Polystyrene	...	10^{21}	...	2.55	2.52	<0.0005	0.0025	24
Polyvinyl chloride	...	...	...	3.2	2.8	0.01	0.006	32
Porcelain	10^{12}	...	5×10^{14}	7	....	...	...	5.7
Pyrex glass	10^{12}	...	10^{11}	5.6	4.9	0.01	0.01	14
Pyranol[7]	...	...	...	5.3	2.7	0.001	0.003	
Quartz[8]	10^{17}	10^{10}	10^{6}	3.8	3.8	0.001	0.0001	8
Steatite	10^{12}	...	...	6.0	5.8	0.002	0.001	24
Sulphur[9]	10^{15}	10^{14}	10^{13}	3.44	3.44	0.0006	0.0007	
Titanium dioxide[10]	10^{12}	...	...	100	90	0.002	0.003	6
Transformer oil[11]	10^{11}	...	...	2.24	2.18	<0.001	0.003	12
Water[12]	5×10^{3}	...	...	....	77	Large	0.15	

[1] The above values are largely from the work of Von Hipple and are for the neighborhood of room temperature.
[2] R.H. = relative humidity.
[3] L.F. = low frequency (10^{2} sec.$^{-1}$).
[4] H.F. = high frequency (3×10^{9} sec.$^{-1}$).
[5] M.F. = medium frequency (10^{3} sec.$^{-1}$).
[6] Small thickness.
[7] Power factor 0.3 at 10^{6} sec.$^{-1}$.
[8] Power factor 0.005 at 10^{6} sec.$^{-1}$.
[9] Power factor 0.0004 at 10^{6} sec.$^{-1}$.
[10] Power factor 0.0003 at 10^{6} sec.$^{-1}$.
[11] Power factor 0.0005 at 10^{6} sec.$^{-1}$.
[12] Power factor 0.03 at 10^{6} sec.$^{-1}$.

ture, pressure, strain, etc., but the atomic theory of the solid state is not sufficiently advanced to yield quantitative information in regard to them. In general, however, an increase in temperature is detrimental to both the mechanical and electrical properties of a dielectric. Thus, since the dielectric power loss increases the temperature and itself increases with increasing field strength a dielectric material tends to fail at high potentials, or high powers. The dielectric loss results either from the motion of charges through the medium or from viscous losses arising from the tendency of permanent molecular dipoles to orient themselves in the electric field. For homogeneous materials the former varies inversely with frequency. The effect of dipole orientation has interesting variations with frequency. As the frequency increases, the dipoles begin to lag more and more behind the field, and their contribution to the dielectric constant decreases. At the same time the power factor first increases, since viscous forces increase with the velocity, then goes through a maximum and finally decreases since the dipoles have negligible motion at high frequencies. As the viscous forces are temperature sensitive the power factor and dielectric constant may exhibit considerable variation with temperature.

FIG. 3.3.—Boundary layer of an electrolytic condenser.

Anisotropic Solids.—There are a number of interesting and important solid dielectrics that have unusual anisotropic properties. One of these is the oxide type of film that forms at the interface between certain metals, such as aluminum, magnesium, or tantalum, when it is the positive electrode in a suitable electrolytic solution. This film has a low resistance to the flow of electricity through it in one direction but below a certain critical voltage presents a high resistance to a flow of current in the opposite sense. The characteristics of this type of surface layer will be considered from the point of view of conduction in Sec. 5.7, but here it is interesting to consider the film as a dielectric separating two conducting media: the metal and the solution. Below the critical voltage, which is determined by the voltage at which the film is formed, it acts as a dielectric but only for one sense of the applied potential. In consequence it cannot be used as a condenser for alternating-potential work but only for a direct potential applied in the proper sense. With this restriction, however, the condenser so formed is of great importance. The film is extremely thin, being of the order

FIG. 3.4.—Schematic representation of the permanent orientation of the molecules in an electret.

of 10^{-5} to 10^{-7} cm. thick. In consequence the capacity per unit area is very high. For a film formed on an aluminum anode (positive electrode) in an ammonium borate solution at 30 volts applied potential the capacity is of the order of 0.2×10^{-6} farad per square centimeter. At 10 times this forming voltage the film is about 10 times as thick, leading to one-tenth the capacity. With films of this type very large capacities may be constructed in a small volume. As an example a condenser employing this principle only 5 cm. on a side may have a capacity as high as 10^{-3} farad. For a very thin film the breakdown potential is low, in general slightly less than the forming voltage. But in common with liquid dielectrics the film has the advantage of being self-healing after breakdown.

There are certain types of dielectrics, of which Carnauba wax is an example, that have unusual electrical properties. If Carnauba wax is melted and allowed to solidify in a strong electric field, it is found that the volume polarization is "frozen in" and remains after the field has been removed. The molecules are free to orient themselves in the liquid state, but owing to the constraints of the solid state retain their orientation over long periods of time. Such substances having a permanent electric moment are known as *electrets*. If an electret is suspended in an electric field, it will orient itself in such a way that the positively charged face is in the direction of the field. It can be cut into smaller blocks, each one retaining its share of the electric moment and exhibiting all the properties of the larger electret. It is evident that a substance must be anisotropic and have a very high resistivity if it is to be capable of becoming an electret.[1]

There are certain classes of crystals with a low degree of symmetry of which tourmaline, tartaric acid, quartz, and Rochelle salt are examples that possess interesting and important electromechanical properties. If mechanical stresses are suitably applied to these crystals, a separation of charge takes place, certain regions becoming positively charged and others negatively charged. The phenomenon is known as the *piezoelectric effect.*[2] In the cases of all the examples given above, except Rochelle salt, a simple compression or extension will result in this surface electrification. In the case of Rochelle salt electrification results from shear (shear will also produce the piezoelectric effect in quartz). The effect is particularly large in the case of Rochelle salt and ammonium dihydrogen phosphate,[3] which has somewhat superior mechanical and electromechanical properties to Rochelle salt. Use is made of it in the construction of electroacoustical devices (see Sec. 14.3). Though the piezo-

[1] GUTMANN, *Rev. Mod. Phys.* **20,** 457 (1948).
[2] CADY, "Piezoelectricity," McGraw-Hill Book Company, Inc., New York, 1946.
[3] MASON, *Phys. Rev.*, **69,** 173 (1946).

electric effect of quartz is small, its mechanical properties are excellent and it is also widely used in piezoelectric devices (also see Sec. 14.3).

The effect has been most carefully studied in the case of quartz, and it is found that when a sample is compressed along the axis of symmetry, no effect is observed; but when compressed perpendicular to this axis, charges appear on the surfaces. The simplest case is illustrated in Fig. 3.5. Assume that if the slab is compressed as shown at the left along an axis perpendicular to that of crystal symmetry, negative charges appear on the upper surface and positive ones on the lower surface. If the crystal is then extended along this axis, the opposite polarity of charges will appear over the two surfaces. Thus the polarity of the charge depends on the sense of the mechanical stress. To a first approximation the amount of charge liberated at one of the surfaces is proportional to the mechanical force applied to it, or the charge per unit area is proportional to the pressure (force per unit area). The piezoelectric constant of quartz, which is the ratio of the charge liberated to the mechanical force, $(\partial q/\partial F)$, is 6.4×10^{-8} esu. per dyne or 2.2×10^{-12} coulomb per newton. The piezoelectric constants for tourmaline and tartaric acid are of the same order of magnitude, while the shear coefficient for Rochelle salt is approximately a thousand times as great.

FIG. 3.5.—Surface charges developed by a simple piezoelectric substance under stress.

The members of certain crystal classes such as tourmaline, sucrose, and tartaric acid may develop charges on certain faces when subject to hydrostatic pressure.[1] Quartz and Rochelle salt do not belong to these crystal classes and hence are not suitable for use in certain types of pressure-measuring instruments.

The converse piezoelectric effect is also observed. If the quartz crystal is placed in an electric field, the crystal either expands or contracts depending on the sense of application of the field. It is the exact converse of the first effect, and the extension or compression of the crystal is proportional to the first power of the potential difference applied between the surfaces.[2] In this respect it obviously differs from the more familiar electrostriction which is exhibited by practically every dielectric. As indicated at (*a*) in Fig. 3.6, a dielectric always tends to increase its linear dimensions in the direction of the field under electrostrictive

[1] LAWSON and MILLER, JR., *Rev. Sci. Instruments*, **7**, 297 (1942).

[2] It may be shown from the principles of conservation of charge and energy that the ratio of the elongation to the potential difference $(\partial l/\partial V)$ is the negative of the piezoelectric constant $(\partial q/\partial F)$.

forces. The positive and negative charges induced on the faces opposite the electrodes tend to distort the material on which they appear in the same sense irrespective of the direction in which the field is applied. In the case of a piezoelectric substance one may think of the crystal as made up of small permanently polarized microscrystalline units that are held in position and orientation by anisotropic body forces. When electrical or mechanical forces are applied to the crystal, the orientation and relative position of all of these microscopic units are slightly changed, resulting in mechanical distortion and the appearance of charges over certain portions of the surface. These crystals also distort anisotropically with change in temperature and the charges so induced are known as *pyroelectric* charges.

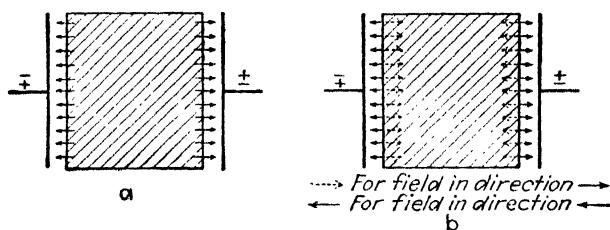



FIG. 3.6.—(*a*) Electrostriction, distortion of a dielectric in a field. (*b*) Simple piezoelectric compression or extension.

3.3. Physical Characteristics of Typical Conductors.—The metals as a class approach most nearly the ideal of a perfect conductor. Very small electric fields will effect a rapid transfer of charge from one region to another. However, this transfer is accompanied by an irreversible transformation of electrical energy into the form of heat. This volume generation of heat, which is known as *joule heating*, is a measure of the resistance presented by the conductor to the flow of current. A metal possesses a typical crystalline structure and is characterized by the presence of conduction electrons which are able to move more or less freely through the ionic crystal lattice. However, the motion of the conduction electrons under the influence of an applied field is at least partially conditioned by the crystal-lattice forces that hold the ionic centers in position. These forces are never completely isotropic and in consequence the crystal lattice presents a greater resistance to electron flow in certain crystal directions than in others. This effect is observed in large single crystals, but it is relatively unimportant in ordinary metallic conductors for these are mosaic structures of microcrystals and the average macroscopic resistance to current flow is the same in any direction.

The ratio of the electric field within a conductor to the rate of flow of charge per unit area in the direction of the field is known as the resistiv-

ity of the substance and is generally written ρ. This ratio is an approximate constant characteristic of the material. Its reciprocal, which is known as the conductivity, is written as σ and either of these parameters can be used to describe the resistance offered by a conductor to the flow of current. Pure silver has the lowest resistivity or highest conductivity of any of the metals at ordinary temperatures. Copper, however, is only slightly inferior to silver in this respect and its relatively low cost makes it the most important metal for electrical purposes. Aluminum has a higher resistivity and a lower tensile strength. However, it has a smaller density which makes it valuable for long conducting spans that must support their own weight or in other services where weight must be kept to a minimum. Though the resistivity of steel is still greater, it is frequently used as an electrical conductor where mechanical strength is an important factor. Mercury is the only metal that is a liquid at ordinary temperatures and this property makes it of great importance in many electrical applications. In addition to these pure metals many alloys have been developed for special purposes. The most important are those of high resistance and high melting point which are used for heating purposes.

The resistivity of a metal is a constant independent of the current density over a very wide range. As an example in the case of gold there is no appreciable change in the resistivity for current densities as high as 10^6 amp. per square centimeter and a change of only a few per cent at 10 times this current. Conductors for which this linear relation exists between the current density and field strength are said to be *linear* or *ohmic* conductors. A simple relation of this type implies that the motion of a conduction electron through a metal may be thought of qualitatively as the motion of a material particle through a viscous medium in which the retarding force is proportional to the velocity. If e and m are respectively the charge and mass of an electron and k is the average effective retarding force per unit velocity, the equation of motion for an electron, say in the x direction, is

$$m\frac{d^2x}{dt^2} + k\frac{dx}{dt} = eE_x$$

When the steady state is reached the acceleration vanishes and the velocity is given as $(e/k)E_x$. Thus, if these are n conduction electrons per unit volume moving with this velocity, the rate of transfer of charge across a unit area perpendicular to the field, which is the current density, is $(ne^2/k)E_x$ and the resistivity is k/ne^2. A discussion of the actual theory of electronic conduction in metals is beyond the scope of this treatment, but the concept of a cloud of conduction electrons drifting through the ionic crystal lattice under the influence of the applied field

and the retarding lattice forces is adequate for visualizing the simpler phenomena of metallic conduction.[1]

The resistivity of a metal depends slightly upon its state of strain which affects the spacing and relative positions of the ionic-lattice centers. This fact is made use of in the *resistance strain gauge*. Such a gauge consists of a thin wire of an alloy, having a resistance insensitive to other factors such as temperature, which is compelled to undergo the same strain as the object being tested. In practice the gauge consists of a long fine wire doubled back and forth in the form of a long close grid, which is cemented to a piece of paper. When this is, in turn, cemented to the member under test, the fractional change in resistance of the wire is a measure of the strain parallel to the grid. The rigidity of the gauge is generally negligible in comparison with that of the member being tested. Free fine wires can be used in a similar manner to measure the relative displacement between two objects to which the ends of the wire are affixed. This principle is employed in the construction of unbonded strain-gauge type of accelerometer. In most metal wires the fractional change in resistance is about twice the longitudinal strain. Strains as small as 10^{-6} can easily be detected, and gauges for dynamic testing up to audio frequencies have been developed.[2]

The most important parameter influencing the resistivity is, however, the temperature. For all pure metals the resistivity increases with rising temperature, but certain alloys such as manganin and constantan have been developed for which the resistivity is practically independent of the temperature over a certain limited range. In the case of carbon the effect is opposite and the resistivity decreases at higher temperatures. The dependence of ρ on T can be expressed over a wide range by a few terms of a power series, *i.e.*,

$$\rho = \rho_1[1 + \alpha(T - T_1) + \beta(T - T_1)^2 + \gamma(T - T_1)^3 + \text{etc.}] \quad (3.2)$$

where ρ is the resistivity at the temperature T and ρ_1 the resistivity at the temperature T_1. The coefficients α, β, γ, etc., decrease rapidly in magnitude and for a range of a few hundred degrees in the neighborhood of room temperature it is sufficient for most practical purposes to neglect all but the first. This coefficient α is generally known as the *temperature coefficient of resistance*. It is an important characteristic of the substance and is listed together with the resistivities in Table IV. At

[1] For an account of the electronic theory of metallic conduction see Hume-Rothery, "The Metallic State," Oxford University Press, New York, 1931; Slater, *Rev. Mod. Phys.*, **6**, 209 (1934); Wilson, "Theory of Metals," Cambridge University Press, London, 1936; Seitz, "The Modern Theory of Solids," McGraw-Hill Book Company, Inc., New York, 1940.

[2] MEYER, *Instruments*, **19**, 136 (1946).

very low temperatures the resistivity of a pure metal becomes very small and practically independent of the temperature. The value of the resistivity approached depends markedly on the presence of impurities and mechanical strains.

In the cases of certain metals, notably lead, tantalum, vanadium, mercury, and tin, the conduction properties change abruptly and the resistivity becomes vanishingly small at a temperature in the neighborhood of a few degrees absolute. Thus in the case of lead the resistivity drops very rapidly at 7.26°K., and a few degrees below this temperature the resistivity is less by a factor of 10^{-12} than at 0°C. For tantalum the transition temperature is 4.38°K., for mercury it is 4.12°K., and for tin, 3.69°K. It was found by Kamerlingh Onnes that a current of several hundred amperes started in a lead ring at 4.2°K. did not decrease by as much as 1 part in 40,000 per hour. This very remarkable phenomenon is known as *superconductivity* and it has been widely studied in recent years in all cryogenic laboratories. Only a limited number of metals become superconductors at the temperatures that have been achieved. But alloys of metals which are not themselves superconductors, such as Au_2Bi and a number of chemical compounds such as CuS, exhibit the phenomenon at sufficiently low temperatures. The lines of current flow in a superconductor are constrained in some way to retain their directions relative to the conductor. That is, if a current is established

TABLE IV

Substance	Resistivity, micro-ohm cm. at 20°C.	Conductivity, mhos/m.	Temperature coefficient (α) per °C.
Aluminum	2.828	3.53×10^7	0.00423
Carbon (graphite)	400–1150		−0.0006 to −0.0012
Constantan	49	2.04×10^6	Negligible 0–100°C.
Copper (annealed standard)	1.7241	5.80×10^7	0.00427
Iron (99.98 per cent pure)	10	10^7	0.0050
Steel (soft)	11.9	8.4×10^6	0.00423
Lead	22	4.55×10^6	0.00411
Manganin	44	2.28×10^6	Negligible 0–100°C.
Mercury	95.783	1.04×10^6	0.00089
Nichrome III (80 per cent Ni, 20 per cent Cr)	103	9.7×10^5	0.00011
Nichrome IV (85 per cent Ni, 15 per cent Cr)	89	1.12×10^6	0.00011
Nickel	7.8	1.28×10^7	0.006
Platinum	9.89	1.01×10^7	0.003
Silver (99.78 per cent pure)	1.629	6.13×10^7	0.0038
Tantalum	15.5	6.45×10^6	0.0033
Tungsten	5.51	1.81×10^7	0.0045
Zinc	5.46	1.84×10^7	0.00402

in a spherical shell in the direction of the circles of latitude, the shell may be rotated about an axis in the equatorial plane and the lines of flow will rotate with the conductor rather than remaining stationary in space. These currents are generally started by magnetic induction (see Sec. 10.2) and it is found that above a certain critical magnetic

field the superconducting properties are lost. But if the superconductor is placed in a magnetic field stronger than this critical value and the field is then reduced, electric currents are induced when the field falls below this value and they will persist for many hours. Their presence is detected by external magnetic effects. A superconductor carrying persistent currents resembles a permanent magnet in certain respects, but the ease with which additional circulating currents may be induced in it makes its properties somewhat different. The theoretical analysis of superconductivity presents many difficulties, but qualitatively a superconductor may be considered as a large diamagnetic atom in which certain of the electrons have the same type of freedom of motion that is possessed by the electron structure of an atom.[1]

3.4. General Theory of Electric Conduction.—The technique that has been employed in the preceding chapters is immediately applicable to the discussion of conduction phenomena. It will be seen that there is a close analogy between the electrostatic field and the state of dynamic equilibrium represented by the steady flow of electric charge. The chief difference is that in the latter case there is a continuous dissipation of electrical energy in the form of heat. It was seen in Sec. 2.1 that the establishment of a charge configuration requires mechanical work, but once the static condition is achieved, there is no further dissipation of energy. Also for perfect conductors and dielectrics the conversion of mechanical into electrical energy is reversible and all the electrical energy stored in the field may be reconverted into the mechanical form. But for real conductors and dielectrics there are irreversible losses at any interconversion process and heat is generated throughout the materials. In the steady flow of electricity through a conductor the electrical forces do work against the crystal forces retarding an electron's motion and this work, which must be supplied by some external source such as a battery or generator, appears in the form of joule heat.

FIG. 3.7.—The rate of flow of charge through the area ds is the product of the current density and the projection of $\mathbf{ds}$ in the direction of flow.

The rate of passage of electric charge per unit area normal to the direction of flow in the neighborhood of any point in a medium is said to be the current density at the point. This is evidently a vector quantity and will be written $\mathbf{i}_v$. The components of this vector along the three Cartesian axes represent the rate of passage of charge per unit area normal to these axes. And in general the rate of flow of charge

[1] For a brief account of these low-temperature phenomena the reader is referred to Jackson, "Low Temperature Physics," Methuen & Co., Ltd., London, 1934; Burton "Superconductivity," University of Toronto Press, 1933.

through an area ds whose normal makes an angle θ with $\mathbf{i}_v$ is $i_v\, ds \cos \theta$ or $\mathbf{i}_v \cdot \mathbf{ds}$, where $\mathbf{ds}$ is the vector of length ds normal to the area. If a hypothetical closed surface is considered, the surface integral of normal outward flow through this surface must be equal to the rate of decrease of charge within it. Calling this charge q, the result would be written

$$\int_s \mathbf{i}_v \cdot \mathbf{ds} = -\frac{\partial q}{\partial t} \tag{3.3}$$

This surface integral may be transformed by the theorem of flux (Appen-

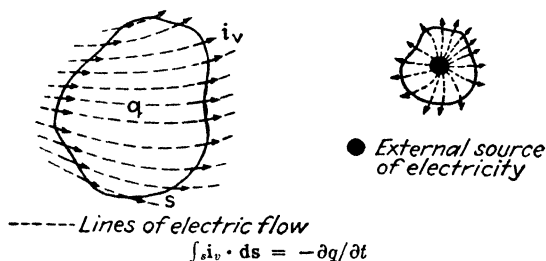



FIG. 3.8.—Illustrations of the surface integral of normal flow and the divergence of the current.

dix D) into an integral taken throughout the volume enclosed by the surface. Since $q = \int_v q_v\, dv$, the equation becomes

$$\int_v \operatorname{div} \mathbf{i}_v\, dv = -\int_v \frac{\partial q_v}{\partial t}\, dv$$

By a similar argument to that used in the derivation of Eq. (1.28) the integrands may be equated, yielding the following differential equation relating the current density and rate of change of charge density at a point:

$$\operatorname{div} \mathbf{i}_v = -\frac{\partial q_v}{\partial t} \tag{3.4}$$

In the steady state that is eventually reached in the case of metallic conduction the charge density is constant (zero in the case of a homogeneous medium) and the right-hand sides of Eqs. (3.3) and (3.4) vanish, yielding

$$\int_s \mathbf{i}_v \cdot \mathbf{ds} = 0 \tag{3.5}$$

and

$$\operatorname{div} \mathbf{i}_v = 0 \tag{3.6}$$

Equation (3.5) shows that there is no net flow through any closed surface as indicated schematically in Fig. 3.8. In the ordinary case of steady flow in a medium there must be at least two areas of contact between the terminals of the source of power maintaining the flow and the medium

itself. Charges enter through one of these areas and leave through the other. These extensions of the terminals of the external battery or generator which make electrical connection with the medium are generally known as *electrodes.* Consider a hypothetical surface that is so drawn as to enclose one electrode as shown at the right in Fig. 3.8. It is evident from Eq. (3.5) that if i is the current entering this surface through the electrode

$$i = \int_{s'} \mathbf{i}_v \cdot \mathbf{ds}$$

where s' is the portion of the surface that does not pass through the electrode but lies in the medium.

In order to proceed any further with the analysis of current flow, it is necessary to assume some functional relation between the current density and the electric field producing it. In the preceding section it was seen that to a very good approximation there is a simple linear relation between these two quantities for metallic conduction. If the medium is not isotropic, the vectors $\mathbf{E}$ and $\mathbf{i}_v$ are not in general in the same direction and the relation between them that is characteristic of the properties of the medium at the point cannot be written in a simple scalar form. However, the most important practical case is that of an isotropic medium in which the current flow is in the direction of the electric field, and there is a simple scalar constant of proportionality between the magnitudes of the vectors. This relation may be written

FIG. 3.9.—(*a*) Current and field vectors in an anisotropic medium. $\mathbf{i}_v$ and $\mathbf{E}$ are not colinear and $\underset{\approx}{\sigma}$ is what is known as a tensor. (*b*) Current and field vectors in an isotropic medium.

$$\mathbf{i}_v = \sigma \mathbf{E} \qquad \text{or} \qquad \mathbf{E} = \rho \mathbf{i}_v \tag{3.7}$$

where σ is the *conductivity* of the medium and ρ is its *resistivity.* Obviously one is the reciprocal of the other. Equation (3.7) is known as *Ohm's law* and a medium in which this relation holds is said to be *ohmic.* If the medium is not homogeneous, σ and ρ are functions of the coordinates, but if the medium is uniform, these parameters are constants independent of position.

When Ohm's law is obeyed, the results of the preceding chapters may be applied very simply to the state of dynamic equilibrium represented by a steady volume distribution of current. Equation (3.6) becomes

$$\text{div } (\sigma \mathbf{E}) = 0 \qquad (3.8$$

and if the medium is homogeneous so that σ is independent of the coordinates, this reduces to div $\mathbf{E} = 0$ or $\nabla^2 V = 0$, which is Laplace's equation. The current density may be derived from the same potential function as the field $\mathbf{E}$ on simply multiplying the latter by the constant σ. The equipotentials are the same whether conduction takes place or not and the lines of current flow coincide with the lines of electric force, regardless of the magnitude of the conductivity. At the boundary surface between two media of different conductivities the lines of force and lines of flow are refracted in a manner analogous to that of the lines of force at the boundary between two dielectrics. Consider a shallow pillbox surface enclosing a portion of the interface. Equation (3.8) implies through the theorem of flux that the surface integral of the normal component of $\sigma\mathbf{E}$ over the surface vanishes. If the box is very shallow, the contribution to the integral that is made by the sides may be neglected in comparison with that made by the faces. If ds is the area of these faces,

FIG. 3.10.—Refraction of lines of flow at the boundary between two conducting media.

$$\sigma_1 E_{1n}\, ds - \sigma_2 E_{2n}\, ds = 0$$

or $\sigma_1 E_{1n} = \sigma_2 E_{2n}$, where the subscripts 1 and 2 refer to the two media on either side of the boundary and the subscript n indicates the normal component of the field. Since the tangential component of $\mathbf{E}$ is the same on both sides of the boundary, this equation implies that the relation between the angles made by the lines of force and flow with the normal to the surface on the two sides is

$$\sigma_1 \cot \theta_1 = \sigma_2 \cot \theta_2 \tag{3.9}$$

as indicated in Fig. 3.10. If the conductivity of one medium is very much greater than that of the other, say $\sigma_1 >> \sigma_2$, then θ_1 is very much larger than θ_2, and the lines of force and flow tend to enter the medium of higher conductivity normally. In the medium of higher conductivity they tend to be parallel to the surface of separation.

The conductivities of different substances differ much more widely than do dielectric constants. In the case for instance of copper electrodes immersed in a conducting solution the lines of flow enter the electrodes practically normally, for their conductivity is greater than that of the solution by a factor of the order of 10^6. There is an even greater factor between the conductivity of even a relatively poor conductor such as a solution and the conductivity of air above it, so the lines of flow are parallel to the exposed surface of a liquid. These facts are made use of

to deduce the forms of the equipotentials and lines of force corresponding to configurations of charged conductors from experimental determinations of the equipotentials and lines of flow in the case of similar configurations of electrodes immersed in a conducting solution. A simple experimental arrangement for obtaining two-dimensional flow patterns is indicated in Fig. 3.11. A conducting solution is contained in a large insulating tray or tank. In this are immersed two parallel electrodes which are connected to the terminals of a battery. The electrodes are shown as circular cylinders and the equipotentials and lines of flow that would be obtained with them are the equipotentials and lines of force

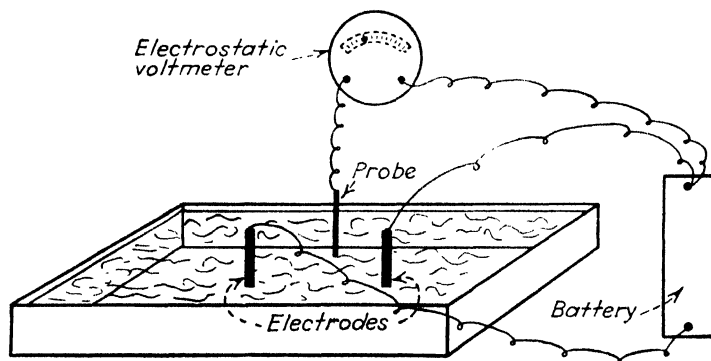



FIG. 3.11.—Tank containing a conducting solution for investigating two-dimensional flow patterns.

in Fig. 1.28. The electrodes, however, may be of any cross section and the technique is particularly useful in the case of geometries that do not lend themselves to calculation. One electrode is connected to one terminal of an electrostatic voltmeter and the other terminal is connected to a probe that may be moved about through the solution parallel to the electrodes. The meter reading then gives the potential difference between the probe and the electrode. If the probe is so moved that the meter reading remains constant, an equipotential curve is traversed. If the probe is attached to one arm of a pantograph, the curve may be traced on paper by the other arm. In this way the family of equipotentials may be drawn and the orthogonal lines of force may be sketched in. The same general technique may also be used in three dimensions if a deep tank is available and all but the extreme point of the probe is covered with a tightly fitting insulating sheath.

When electrodes are immersed in an ohmic medium, it is evident that the total current entering at one and leaving at the other is proportional to the potential difference between the electrodes. The ratio of the potential difference to the current is said to be the *resistance*

of the medium between those electrodes. The relation between the total current i and the potential difference V is written

$$V = Ri \tag{3.10}$$

where R is the resistance. Equation (3.10) is an alternative statement of Ohm's law. The unit of resistance is the *ohm, i.e.*, if a current of 1 coulomb per second or 1 ampere flows between the electrodes when their potential difference is 1 volt, the resistance between the electrodes is 1 ohm. The resistivity, being of the dimensions of field strength divided by current density, would be in ohm-meters in the units here adopted. In Table IV it is given in micro-ohm-centimeters; to convert these entries to ohm-meters they must be multiplied by 10^{-8}. The reciprocal of the resistance is called the *conductance* and its unit, the *mho,* comes from a reversal of the order of the letters in ohm. The conductivity, which is the reciprocal of the resistivity, is generally given in mhos per centimeter or mhos per meter as in Table IV.

There is a very useful relation between the resistance presented by two electrodes and the electrostatic capacity that would exist between them if all the intervening material were removed. The current that flows from an electrode in an ohmic medium is given by

$$i = \int_{s'} \mathbf{i}_v \cdot \mathbf{ds} = \sigma \int_{s'} \mathbf{E} \cdot \mathbf{ds}$$

where s' is a surface enclosing the electrode except for the area through which current enters it. The contribution of this excluded area may generally be made negligible in the electrostatic problem and the integral $\int_s \mathbf{E} \cdot \mathbf{ds}$ over a completely enclosing area is equal to the charge on the electrode, considered as an isolated conductor, divided by κ_0 (Gauss's theorem). Considering the s and s' integrals as identical the equation becomes

$$i = \frac{\sigma}{\kappa_0} q$$

where q is the charge on the electrode that would produce the field $\mathbf{E}$. The electrostatic capacity between the two electrodes considered as the surfaces of a condenser is $C_0 = q/V$, where V is the potential difference between them. Therefore

$$i = \frac{\sigma}{\kappa_0} C_0 V$$

or R, which is the ratio of V to i, is given by

$$R = \frac{\kappa_0}{\sigma C_0} \tag{3.11}$$

Equation (3.11) is very useful in giving the resistance between two electrodes of high conductivity and calculable capacity immersed in a medium of known conductivity. It may also be used to calculate the conductivity if the resistance is measured between electrodes of known geometry.

Consider, for example, a rectangular block of material of conductivity σ bounded on two opposite faces by electrodes in the form of plates of very great conductivity connected to the terminals of a battery. These plates may be considered as portions of an infinite parallel-plate condenser. If their area is A and separation d their capacity, assuming that the lines of force run uniformly between them, is $\kappa_0 A/d$ by Eq. (1.24′). This approximation would be poor for the geometry of Fig. 3.12 in the case of a condenser, but it is very good for calculating the resistance presented by the block as the lines of flow are constrained to remain in the block and will be approximately uniformly distributed over any cross section. By Eq. (3.11) the resistance presented by the block is

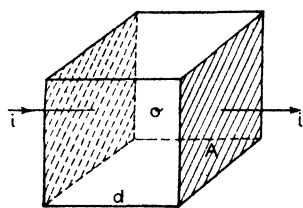



FIG. 3.12.—Plates of area A and of infinite conductivity on opposite faces of a block of material of conductivity σ.

$$R = \frac{d}{\sigma A} \tag{3.12}$$

Thus the resistance presented by a block of unit area and unit length is equal to the reciprocal of the conductivity or to the resistivity. Equation (3.12) may be used to calculate the resistance presented by a long conductor of uniform cross section and known conductivity. If the cross section is not uniform the calculations cannot in general be performed exactly. If the lines of flow were always perpendicular to the cross sections, the effective resistance of a conductor of variable section could be obtained by applying Eq. (3.12) to each infinitesimal segment, yielding

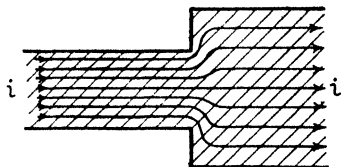



FIG. 3.13.—Section through a conductor indicating the behavior of the lines of flow in the neighborhood of a change in cross section.

$$R = \frac{1}{\sigma}\int_l \frac{dl}{A}$$

where l is the coordinate in the direction of the length. However, the lines of flow tend to follow the contours of the conductor as shown schematically in Fig. 3.13, and this integral gives merely a lower limit to the resistance.

Another important case is that of the two coaxial cylindrical electrodes depicted in Fig. 3.14. If these are lowered into a liquid so that they rest on the flat bottom of the container and extend up above the surface or if

the region between them is filled with a medium of much lower conductivity than the electrodes, the lines of flow between them are radial in close analogy with the lines of force between the plates of an infinite cylindrical condenser. By Eq. (1.23) the capacity of a cylindrical condenser of these dimensions is

$$C_0 = \frac{2\pi\kappa_0 l}{\log_e (b/a)}$$

Therefore by Eq. (3.11) the resistance presented by the material between the electrodes is

$$R = \frac{\log_e (b/a)}{2\pi l \sigma}$$

FIG. 3.14.—Resistance presented to the flow of current by a medium of conductivity σ between two coaxial cylindrical surfaces of infinite conductivity.

This is a very convenient geometry for measuring the conductivity of a fluid. A configuration that is of interest in connection with electrical circuits that involve the ground as a return path is that of two spherical electrodes which are a great distance apart in comparison with their linear dimensions. The electrostatic capacity between two such spheres is given approximately by Eq. (1.32) as

$$C_0 = 4\pi\kappa_0\left(\frac{1}{a} + \frac{1}{b} - \frac{2}{l}\right)^{-1}$$

where a and b are the radii of the spheres and l is their separation. Thus

FIG. 3.15.—Resistance of the earth between two hemispherical electrodes.

the resistance between them if they were embedded in an infinite ohmic medium would be

$$R = \frac{1}{4\pi\sigma}\left(\frac{1}{a} + \frac{1}{b} - \frac{2}{l}\right)$$

If two hemispherical electrodes are sunk below the surface of the earth as shown in Fig. 3.15 and if l is much greater than a or b, but is small in comparison with the radius of the earth so that the surface of the latter may be considered as plane, the resistance between them is twice that given above, assuming that the conductivity of the earth is uniform. This latter assumption is rarely justified except to a very crude approxi-

mation. However, the expression for R shows that the resistance between them is practically independent of their separation if l is large. If the hemispheres are of the same radius, the resistance between them to this approximation is $1/\pi\sigma a$. It is evident that the conductivity of the earth in the immediate neighborhood of the electrodes is of greatest importance and this may generally be increased by moistening the earth around them. The inverse proportionality between the resistance and the linear dimensions applies approximately to electrodes of any shape.

Equation (3.11) may also be used to obtain the electrostatic capacity between two electrodes if the latter cannot be calculated or measured

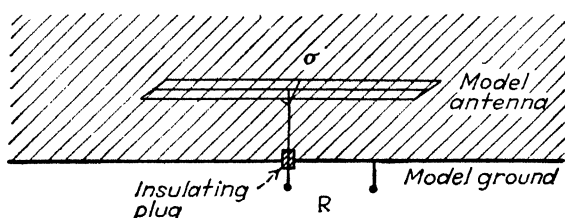
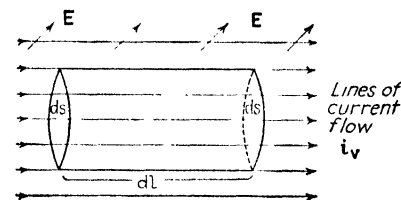



FIG. 3.16.—Representative scale model for measuring the resistance between two electrodes for the purpose of calculating the electrostatic capacity between them.

directly. If the electrodes are immersed in a homogeneous ohmic material and all bounding surfaces except those of the electrodes themselves are at a very great distance, a measurement of the resistance between them permits a calculation of their capacity. Frequently the dimensions of the actual electrodes are too large for this type of measurement and in this case a scale model may be constructed. An example of a radio antenna above a perfectly conducting ground is shown in Fig. 3.16. It was seen in Sec. 1.5 that the capacity between geometrically similar electrodes is proportional to the linear dimensions. Therefore the capacity calculated for the model from Eq. (3.11) must be multiplied by the scale factor between the actual electrodes and that of the model.

Generation of Heat.—The most important characteristic associated with the conduction of electricity and differentiating it from the phenomena of electrostatics is the joule heating that takes place within the conductor. The continuous conversion of electrical energy into heat by this process represents the most important loss of available energy in practically all electrical devices. Incidentally, the heat thus generated raises the temperature of the conductor and alters its resistance, as indicated by Eq. (3.2). If a charge q is transferred between two electrodes at a potential difference V, an amount of energy qV must be expended. This may be extracted from a previously existing charge configuration or it may be supplied by some external source such as a battery or generator. As i represents the rate of transfer of charge

between two electrodes, the product iV is the rate at which the electric forces perform work or the power that is expended

$$P = iV \tag{3.13}$$

In the case of conduction through an ohmic material this equation may be combined with Eq. (3.10) and the power written in the alternative forms

$$P = iV = i^2R = \frac{V^2}{R} \tag{3.14}$$

In the units here adopted qV is in joules and the power is in joules per second or *watts*.

The process taking place in the interior of a conductor may be considered in more detail by the aid of Fig. 3.17. The charge transferred through the area ds perpendicular to the current density $\mathbf{i}_v$ is $i_v\, ds$ per second. In the general anisotropic medium the field $\mathbf{E}$ is not parallel to $\mathbf{i}_v$, but if the angle between these vectors is θ, the potential difference between the two areas ds a distance dl apart is $E \cos\theta\, dl$. Therefore the rate at which the electrical forces are performing work that is being transformed into heat in the volume $ds\, dl$ is $i_v\, ds\, E \cos\theta\, dl$. Therefore the rate of generation of heat in this volume element is $\mathbf{E} \cdot \mathbf{i}_v\, dv$, or the rate of generation of heat per unit volume is

FIG. 3.17.—Rate of working of the electrical forces per unit volume.

$$P_v = \mathbf{E} \cdot \mathbf{i}_v \tag{3.15}$$

If the medium is ohmic with a conductivity σ or resistivity ρ, Eq. (3.15) may be written

$$P_v = Ei_v = \sigma E^2 = \rho i_v^2 \tag{3.16}$$

This equation gives the rate of generation of heat per unit volume in terms of either the electric field or the current density. The expression applies, of course, whether the medium is homogeneous or not; in the latter case σ and ρ are functions of the coordinates.

Poor Conductors.—If the conductivity is low as in the case of imperfect dielectrics, it is necessary to take the effect of the dielectric constant of the medium into account. In this discussion the dielectric and

conduction properties will be assumed to be linear and isotropic, hence in the equilibrium condition Eqs. (2.24) and (3.8) apply:

$$\text{div } (\kappa \mathbf{E}) = \frac{q_v}{\kappa_0}$$
$$\text{div } (\sigma \mathbf{E}) = 0$$

Performing the indicated differentiations, allowing for the possibility that κ and σ are functions of the coordinates

$$\kappa \text{ div } \mathbf{E} + \mathbf{E} \cdot \mathbf{grad}\, \kappa = \frac{q_v}{\kappa_0}$$
$$\sigma \text{ div } \mathbf{E} + \mathbf{E} \cdot \mathbf{grad}\, \sigma = 0$$

Eliminating the divergence term between these equations, they yield

$$\sigma \mathbf{E} \cdot \left(\frac{\mathbf{grad}\, \kappa}{\sigma} - \frac{\kappa}{\sigma^2} \mathbf{grad}\, \sigma \right) = \frac{q_v}{\kappa_0}$$

The quantity $\sigma \mathbf{E}$ is the current density and the terms in brackets are together equal to **grad** (κ/σ), as may be seen by expansion. Therefore a current density $\mathbf{i}_v$ in an inhomogeneous conducting medium implies the existence of a charge density q_v given by

$$q_v = \kappa_0 \, \mathbf{i}_v \cdot \mathbf{grad} \frac{\kappa}{\sigma} \tag{3.17}$$

If the inhomogeneity is such that κ/σ is constant, the charge density vanishes, but this condition is seldom fulfilled in practice.

For good conductors the electric field is small for ordinary currents and this volume density of charge is generally negligible. In the case of an imperfect dielectric, however, σ is small and the volume charges that arise when a field is applied and leakage current flows through the substance are appreciable and are at least partially responsible for the phenomenon of *dielectric absorption*. This is the occurrence of free charges throughout the dielectric and particularly at boundaries between two dielectric media. It is of practical importance in the case of large condensers operating at high potentials and employing the ordinary types of commercial dielectrics. The condition at the boundary between two homogeneous media can be obtained from Eq. (3.17) or from Eq. (2.28). The latter may be written

$$\kappa_1 E_{1n} - \kappa_2 E_{2n} = \frac{q_s}{\kappa_0}$$

and since the normal components of the field may be written in terms of the conductivities and the normal component of the current density

$$i_{vn} = \sigma_1 E_{1n} = \sigma_2 E_{2n}$$

Thus the surface charge density at the interface becomes

$$q_s = \kappa_0 i_{vn}\left(\frac{\kappa_1}{\sigma_1} - \frac{\kappa_2}{\sigma_2}\right)$$

Though these volume and surface charges associated with imperfect dielectrics are generally small, their effects may be of importance in large condensers employing certain types of inhomogeneous dielectrics. If a charged condenser is discharged by a spark between its terminals and the plates then insulated from one another the diffusion of the charges out of the dielectric to the plates may give rise to a very appreciable residual charge which will be retained for a considerable time if the

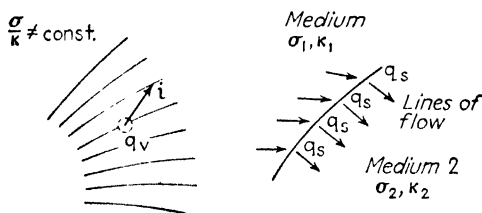



FIG. 3.18.—(a) Volume charge density in an inhomogeneous current-carrying medium. (b) Colinearity of **D** and i in the two media on either side of a boundary implies a charge distribution at the interface.

conductivity of the dielectric is low. The *relaxation time* of a dielectric is a measure of the time necessary for this diffusion and redistribution of charge that takes place when a field is applied or removed. Assuming for the purpose of the immediate argument that the dielectric is approximately homogeneous, Eq. (3.4) may be used in the form

$$\sigma \operatorname{div} \mathbf{E} = -\frac{\partial q_v}{\partial t}$$

and since for a homogeneous dielectric div $\mathbf{E} = q_v/\kappa\kappa_0$, the equation becomes

$$\frac{\partial q_v}{\partial t} = -\frac{\sigma}{\kappa\kappa_0} q_v$$

This equation may be integrated immediately to yield

$$q_v = q_{v0} e^{-\frac{\sigma t}{\kappa\kappa_0}}$$

where q_{v0} is the volume density of charge at the time $t = 0$. From this equation it is evident that the time which elapses during the decrease of the charge density in a region to $1/e$ of its original value is $\kappa_0\kappa/\sigma$. If the conductivity is large as in the case of substances in Table IV or even for most liquid dielectrics in Table II, this relaxation time is extremely

small. For petroleum oil, however, this time is of the order of 2 sec. and for a substance such as sulphur or ceresin it is of the order of days.

Problems

1. The pressure of the air between the plates of a condenser is changed from 1 to 10 atm. Show that the percentage change in capacity is 0.53 per cent. Show that the fractional change in capacity of an air condenser at normal pressure if the temperature is changed from T_1 to T_2 is $0.162\left(\frac{1}{T_1} - \frac{1}{T_2}\right)$.

2. An electret is suspended so that its electric moment is horizontal by a fiber of negligible restoring torque in a horizontal electric field E. Show that the period of small oscillation of the block is $2\pi\sqrt{I/pE}$, where p is the magnitude of its electric moment and I its mechanical moment of inertia.

3. A quartz crystal slab, cut so as to contain the axis of symmetry, is placed between two copper plates that are connected to the terminals of an electrometer. If the sensitivity of the instrument is 2,000 divisions per volt and a deflection of 200 divisions is observed when a mass of 2 kg. is laid on the crystal, show that the effective capacity of the system is 430 $\mu\mu$f.

4. Show that a tenuous suspension of metal spheres in a medium of unit dielectric constant will exhibit a susceptibility equal to $3f$, where f is the fraction of the volume occupied by the spheres.

5. Show that the average velocity of drift of the conduction electrons in copper is 0.736 mm. per second for a current density of 10 amp. per square millimeter if one conduction electron is assumed per atom. (Number of copper atoms per cubic centimeter $= 8.5 \times 10^{22}$.) Show that the mean retarding force per unit velocity exerted by the copper lattice on a conduction electron is 3.75×10^{-14} dynes per centimeter per second.

6. The bottom and sides of a cubic vessel 1 cc. in capacity are of glass and the two ends are of copper. Show that if the vessel is filled with pure water and a potential difference of 100 volts applied between the two copper faces, the temperature of the water will rise at the rate of 4.8×10^{-3} °C. per second. Neglect the loss of heat by conduction and take the calorie as 4.18 joules.

7. The air space between the two shells of a spherical condenser is filled with petroleum oil having a dielectric strength of 50 kv. per millimeter and a dielectric constant of 2.1. Show that the charge the condenser can hold is increased by the factor 35.

8. The capacity of a condenser with petroleum oil for the dielectric is 10 μf. Show that if a potential difference of 500 volts is applied between its terminals, the leakage current will be 2.55 ma. and that heat will be generated within it at the rate of 1.27 watts.

9. A long copper wire of radius a runs through a deep lake at a height h above the plane bottom. Assuming the bottom to be a good conductor, show that the resistance per unit length between it and the wire is

$$\frac{\cosh^{-1}(h/a)}{2\pi\sigma}$$

where σ is the conductivity of the lake water.

10. Two copper rods 1 cm. in diameter and 10 cm. between centers enter a large graphite disk normally and near its center. If the graphite disk is 1 cm. thick and has a conductivity of 1.5×10^5 mhos per meter, show that the resistance between the copper electrodes is 6.36×10^{-4} ohm.

11. A copper sphere 10 cm. in diameter is lowered centrally into a hemispherical copper bowl 20 cm. in diameter that contains water with a conductivity of 10^{-3} mho per meter. Show that the resistance between the electrodes when the spheres are concentric is 1,590 ohms.

12. A spherical electrode of radius a is placed in a medium of conductivity σ at a distance d from a large plane slab of great conductivity. Show that the resistance to the flow of current from the sphere to the slab is

$$\frac{1}{4\pi\sigma}\left(\frac{1}{a} - \frac{1}{2d}\right) \quad \text{if} \quad a < < d$$

13. A condenser employing petroleum oil as the dielectric is charged to a potential V. If the battery is then disconnected, show that the potential difference between the electrodes will fall to 1 per cent of its initial value in 8.95 sec.

14. Show that the time in which the charge on two condenser plates will fall to $1/e$ of its initial value is equal to the product of the capacity of the condenser and the leakage resistance between its terminals.

15. If the field tending to polarize a hydrogen molecule is 1 volt per 10^{-8} cm., show that the electric moment induced is that equivalent to a positive and negative charge equal in magnitude to that of the electron 5.4×10^{-10} cm. apart.

16. Show by direct integration that when a condenser is charged by a battery, the amount of joule heat developed in the circuit is equal to the final electrostatic energy of the condenser.

17. A system of electrodes is imbedded in a conducting medium. They are maintained at the fixed potentials $V_1, V_2, \ldots, V_n$ and the currents leaving the electrodes are $I_1, I_2, \ldots, I_n$. Show that the total rate of generation of joule heat in the medium is

$$Q = \sum_{i=1}^{i=n} V_i I_i$$

18. Two large thin slabs of dielectric material are pressed face to face, and against the outer surfaces are placed conducting plates to form a parallel-plate condenser of separation small in comparison with its linear dimensions. The thicknesses of the two dielectric slabs are a and b, and their dielectric constants and conductivities κ_1, σ_1 and κ_2, σ_2, respectively. A potential difference V is then maintained between the condenser plates. When the steady state is established, find the potential of the dielectric interface with respect to one of the plates and also the surface-charge density at the dielectric interface.

19. Referring to the discussion of two-dimensional electrostatic problems in Sec. 1.8, show that the resistance presented to the flow of current between two terminals which extend through a thin infinite sheet of material of conductivity σ and thickness τ is

$$R = \frac{(V_2 - V_1)}{\sigma\tau(U_2 - U_1)} = \frac{V_2 - V_1}{2\pi\sigma\tau}$$

where the V's as in Sec. 1.8 are the equipotential parameters defining the exposed surfaces of the terminals.

20. Two sets of semicircular clamp terminals are affixed to the edge of a thin circular disk of conductivity σ and thickness τ. They are of such a nature that the center of the circular area of radius a is at the periphery of the disk and the separation

of these centers is s. Show that the resistance R presented to the flow of current between the terminals is

$$R = \frac{2 \cosh^{-1} (s/2a)}{\pi \sigma \tau}$$

21. A very large thin rectangular sheet of material having a conductivity σ has electrodes along opposite edges that maintain an electric field E_0 or supply a current density E_0/σ per unit length to the sheet. A small circular hole of radius a is made near the center of the sheet, which has a negligible effect on the conditions at the electrodes. Show that the subsequent potential distribution in hole and sheet are given, respectively, by

$$V_h = -2E_0 r \cos \theta$$
$$V_s = -E_0 r \cos \theta \left(1 + \frac{a^2}{r^2}\right)$$

where r and θ are polar coordinates about the center of the hole. Show also that the rate of production of heat per unit volume of the sheet is f times that at the same point before the hole was made where

$$f = \left(1 - \frac{2a^2}{r^2} \cos 2\theta + \frac{a^4}{r^4}\right)$$

Show by integration that the total rate of generation of heat is unaltered by the presence of the hole.

CHAPTER IV

DIRECT-CURRENT CIRCUITS

4.1. Introduction.—The electrical circuit most commonly encountered is the type that is composed of sources of electromotive force which are connected together by long, thin metallic filaments or wires. An *electromotive force*, generally abbreviated emf., is any agency which, when present in a conducting circuit, induces a flow of current. The motion of conduction electrons through a wire implies the existence of an electric field in the wire and hence a potential difference between its terminals. This potential difference is produced by the source of emf. The magnitude of the emf. produced by a device is defined as the potential difference that exists between its terminals when no current is flowing through it, *i.e.*, it is the static or open-circuit potential difference. It can be measured by any of the electrostatic instruments that have been previously described, for these instruments satisfy the necessary condition and draw no current. Electromotive forces can also be compared with one another by means of the potentiometer type of circuit which will be described later in this chapter, for this also satisfies the necessary condition and draws no current from the device when in the proper adjustment.

A source of emf. is also a source of power. For, if it induces the flow of a current i through a portion of the circuit of resistance R, heat is generated in that portion at the rate i^2R and this must represent the energy abstracted per unit time by this portion of the circuit from the source of emf. Thus a source of emf. is essentially a device for converting energy into an electrical form. The electrostatic generator and rotating electromagnetic machinery convert mechanical into electrical energy. The various types of voltaic cells, which in combination form batteries, are essentially devices for transforming chemical energy into the electrical form. Likewise, thermocouples and photovoltaic cells, which produce emfs. under the influence of temperature differences or in the presence of electromagnetic radiation, may be thought of merely as agents for the conversion of thermal and radiant energy into electrical energy. In the external circuit the energy may be converted back into heat, or into mechanical energy with a motor, or into chemical energy as in charging a storage battery. These interconversion processes are of course not perfectly efficient and a certain amount of energy is always lost or becomes unavailable in the form of heat. Even electrical circuits designed for

the production of heat are not 100 per cent efficient, for the heat is not all generated at the place where it is required. The various devices mentioned above will be discussed in later chapters. Here we will be principally concerned with the circuit external to the source of emf. The source will generally be assumed to be a chemical cell or battery and will be represented schematically by the symbol $\mathcal{E}$, as in Fig. 4.2.

The greater portion of an ordinary electric circuit is made up of metallic wires that are very long in comparison with their cross-sectional dimensions. Thus over the greater part of their length the lines of current flow are parallel to the boundaries. To a very good approximation the effects at the ends can be neglected and the resistance of a wire of length l and cross section A composed of a metal of resistivity ρ is $\rho l/A$. In the case of copper, which is the metal most frequently used because of its high conductivity and low cost, $R = 1.724 \times 10^{-8} l/A$ ohms, where l is in meters and A in square meters, or $R = 1.724 \times 10^{-6} l/A$, where l is in centimeters and A in square centimeters. Frequently, however, a portion of a circuit is not of such a form that its resistance can be readily calculated or the length and cross section of the wire are not known. In that case the resistance can be determined by measuring the current through the circuit and the potential difference across the required portion and applying Ohm's law in the form $R = V/i$. R is a constant of the circuit for truly ohmic materials and as such it is a quantity of great importance. If R is known, Ohm's law may be used to determine V from a measurement of i or vice versa. An absolute measurement of resistance can in principle be made by measuring the rate of generation of heat. The dissipation of power or rate of generation of heat in a resistance is equal to Vi, where i is the current flowing through it and V is the potential difference between its terminals. From Ohm's law $Vi = V^2/R = P$. V can be measured with an absolute electrometer and P, the heat generated per unit time, can be measured with a calorimeter; thus R can be determined. However, these measurements cannot be performed with an accuracy that is satisfactory for resistance standards and a description of the methods that are actually employed for making absolute determinations of resistance will have to be postponed till Sec. 10.4. The realization of a satisfactory standard of resistance, the ohm, will here be assumed. Methods for determining the value of an unknown resistance by comparison with a standard, which are independent of a knowledge of either i or V, will be described later in this chapter.

4.2. Fundamental Direct-current Circuit Analysis.—The discussion in this chapter is limited to the case in which the emfs. and currents flowing in the circuits are constant and do not vary with the time. Such currents are known as direct currents. The electrons move through

the ionic lattice of the conductor under the influence of the electric field. In the simple case of a single closed conducting circuit, as shown in Fig. 4.1, the number of electrons passing through a cross section of the conductor normal to the lines of flow is the same at any point of the circuit. Otherwise the number of electrons in a region would either increase or decrease with the time, which is contrary to our hypothesis. The circuit of Fig. 4.1 represents a voltaic cell with its terminals connected by a conducting wire of length l_1, cross-sectional area A_1, and resistivity ρ. The emfs. developed at the two plates of the cell are assumed to be $\mathcal{E}_1$ and $\mathcal{E}_2$, though it is only their algebraic sum which can be measured. The resistance of the solution including the boundary layers at the electrodes is assumed to be R_s. Though the origin of the electromotive force, say $\mathcal{E}_1$, is not electrostatic but resides in the chemical nature of the

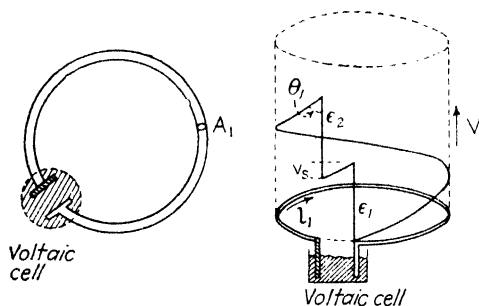



FIG. 4.1.—Potential variation in a closed electric circuit.

bounding layer between the metallic electrode and the solution, it is measured, at least in theory, by the potential difference which it establishes between the two media on either side of the boundary. Thus as far as the circuit exclusive of this boundary is concerned, it may effectively be replaced by an electrostatic potential difference of magnitude $\mathcal{E}_1$. The same remarks apply to the emf., $\mathcal{E}_2$, at the second boundary layer. The general electrostatic theorem that $\oint \mathbf{E} \cdot d\mathbf{l} = 0$ may then be applied. Graphically an ordinate may be erected at every point of the circuit proportional to the potential difference between that point and a fiducial point at one of the plates of the cell. From the general electrostatic theorem this curve must be closed and the algebraic sum of the emfs. must equal the potential drop in the cell plus that in the external circuit, *i.e.*,

$$\mathcal{E}_1 + \mathcal{E}_2 = V_s + V_e$$

V_s can be written as iR_s and V_e as $i\rho l_1/A_1$ or iR_e. Writing the left side simply as $\mathcal{E}$, the open-circuit potential difference between the electrodes

$$\mathcal{E} = iR_s + iR_e$$

The emf. applied to the *external* circuit resulting in the current i is not $\mathcal{E}$

but $\mathcal{E} - iR_s$. Only if the internal resistance of the source of emf. is negligible is the effective emf. applied to the external circuit a constant independent of i.

In drawing a schematic electric circuit, a resistance is indicated by a zigzag line, as illustrated in Fig. 4.2. Also the internal resistance associated with the source of emf. is generally drawn in explicitly next to the symbol for emf., as is also illustrated in that figure. The left-hand portion of this figure represents a simple circuit in which the current is the same throughout. The emfs. are seen to be applied in opposite senses, hence their algebraic sum is $\mathcal{E}_1 - \mathcal{E}_2$. $\mathcal{E}_1$ is chosen as positive to agree with the sense of current flow there indicated, for a positive charge moves through the circuit toward the negative terminal of the source of emf. All the resistance is concentrated in the zigzag lines of such a figure. The straight lines represent ideal resistanceless connections.

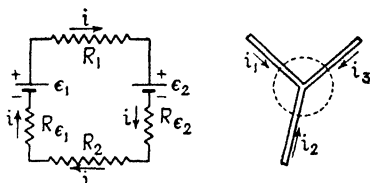



FIG. 4.2.—Illustration of Kirchhoff's laws.

The potential difference between the terminals of a resistance carrying a current is frequently referred to as the "iR drop." A more general circuit would involve the meeting of more than two resistances at a common point. Such a point is known as a *branch point* or *junction*. If an imaginary closed surface is drawn about such a junction, as illustrated at the right in Fig. 4.2, the application of the equation $\int \mathbf{i}_v \cdot \mathbf{ds} = 0$ shows that the algebraic sum of all the currents flowing to a junction is zero. Furthermore, if branch points occur, the current is not in general the same through the different resistances composing the network of branches. Both the resistance R and the current i flowing through it vary from branch to branch. Consider one conducting circuit which forms part of a more complicated network. If it is composed of a resistance R_1 carrying a current i_1, a resistance R_2 carrying a current i_2, and a resistance R_3 carrying a current i_3, and an emf. $\mathcal{E}$ is encountered on traversing the circuit, the equation stating that the sum of the iR drops is equal to the circuit emf. is

$$\mathcal{E} = i_1R_1 + i_2R_2 + i_3R_3$$

Or, in general, if there are n resistances which make up a closed path that may be labeled m through a network of conductors and if $\mathcal{E}_m$ is the total emf. encountered in traversing the path

$$\mathcal{E}_m = \sum_{k=1}^{k=n} i_kR_k \tag{4.1}$$

And at any junction where, let us say, l currents meet, the condition is

$$\sum_{j=1}^{j=l} i_j = 0 \tag{4.2}$$

These two equations are known as *Kirchhoff's laws*. All of the circuit theory of steady currents is based upon these laws and the rest of this chapter will be largely devoted to their application in particular instances.

The simplest circuit is that of a number of resistance in *series*, for in this case there is but one closed path that can be traversed through the conductors. If a potential difference V is applied to the resistances

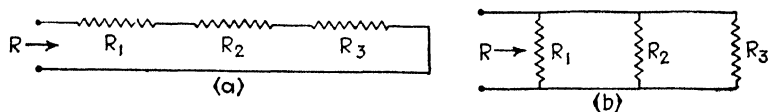



FIG. 4.3.—Elementary resistance configurations. (*a*) Resistances in series. (*b*) Resistances in parallel.

R_1, R_2, and R_3 in series, *i.e.*, connected end to end, the fundamental circuit theorem yields

$$V = V_1 + V_2 + V_3$$

The V's with subscripts represent the potential drops across the corresponding resistances, and $V_1 = iR_1$, $V_2 = iR_2$, and $V_3 = iR_3$; hence

$$V = i(R_1 + R_2 + R)_2 = iR$$

where R is written for the effective resistance presented by the two terminals of the chain of resistances. Thus the effective resistance presented to the rest of the circuit by a chain of resistances in series is the sum of the separate resistances forming the chain. The effective resistance presented by a group of resistances is frequently referred to as the *input resistance* of the portion of the circuit which they form. If a group of resistances are arranged in *shunt* or *parallel* as shown in Fig. 4.3, *b*, the effective resistance which they present to the rest of the circuit can also be expressed very simply by means of Kirchhoff's laws. If V is the potential difference appearing across their terminals, the current through resistance R_1 is V/R_1, that through R_2 is V/R_2, and so forth. The sum of these currents is equal to the total current flowing to the group, which is V over their effective resistance. Hence

$$\frac{V}{R} = i = i_1 + i_2 + i_3 = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3}$$

Thus the effective resistances presented by these two simple configurations are

$$R = R_1 + R_2 + R_3 + \cdots \qquad \text{(series)}$$

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots \qquad \text{(parallel)}$$

In terms of the *conductance* G of the circuit, which is defined as the reciprocal of the resistance, the second equation becomes

$$G = G_1 + G_2 + G_3 \qquad \text{(parallel)}$$

These expressions are immediately applicable in many instances and often more complex circuits can be analyzed by inspection into groups of series and parallel elements. In this way it is possible to determine the effective resistance presented by these circuits and to obtain the current flowing to them if the applied emf. is known. The currents flowing in the rest of the branches of the network can then be determined by the further application of these simple equations.

4.3. General Circuit Theorems.—In general it is not possible to analyze a complicated circuit by inspection into its component series and parallel elements. For a complete solution Eqs. (4.1) and (4.2) must be applied systematically to the network of conductors and the potential differences between the junctions or the currents in the branches obtained. There are two general methods of approach known as *shunt analysis* and *series analysis*, respectively. In shunt analysis the potentials of the junctions are assigned a series of values, and the equations of the network are set up in terms of these as the variables. The current through a *branch* that joins any two junctions together is given by

$$V_l - V_i = i_{il}R'_{il} - \mathcal{E}_{il} \tag{4.3}$$

where $V_l - V_i$ is the difference in potential between the junctions, i_{il} and R'_{il} are the current from i to l and resistance in the branch il, and $\mathcal{E}_{il}$ is any emf. that may be in the branch with its sign chosen in accordance with the convention of Sec. 4.2. As the branch currents are completely determined by the potential differences between junctions, it is clear that the assignment of a potential to each junction is redundant. An arbitrary junction can always be the zero of potential; and hence if J is the number of junctions in the network, the number of independent V's is $(J - 1)$. The equations of the form of Eq. (4.1) are automatically satisfied by the choice of potentials. The equations of the form of Eq. (4.2), stating that the sum of the currents to each of these junctions is zero, are the network equations but are not all independent for a closed network. If the junctions are considered serially, the sum of the currents to the last junction vanishes automatically because of the conditions on the currents leaving all other junctions of the network that are connected to it. Thus there are only $(J - 1)$ equations, and these just suffice to determine the V's and hence by Eq. (4.3) the branch currents. In series analysis a set of closed conducting paths through the branches of the network are chosen, and circulating currents in these paths, which

are known as *meshes*, are assigned values. As these paths are closed, the equations of the form of Eq. (4.2) are automatically satisfied. The equations of the form of Eq. (4.1) are the network equations, which are not all independent for the totality of meshes that can be chosen. Each branch must be included in a mesh at least once in order that all the branch currents may be determined. As there are, say, B branch currents to be eventually determined and as Eqs. (4.2) supply $(J - 1)$ equations of condition, the minimum number of mesh currents that must be chosen for the variables in the equations which is equal to the number of meshes M is given by $M = B - (J - 1)$. As $(J - 1) < M$, except for $J = 3$ shunt analysis requires fewer equations than series analyses; however, each method has its advantages for particular circuits.[1] If the potential differences between certain pairs of junctions are of primary interest, these come immediately out of a shunt analysis provided the potential zero is chosen at one of the junctions involved. If a particular branch current is desired, this comes immediately out of a series analysis where but one mesh traverses that branch. A series analysis is particularly adapted to taking account of additional emf's. that are added later and a shunt analysis can readily be altered to take account of additional external currents that may be introduced later at junctions.

Shunt analysis proceeds by writing the equations of the form of Eq. (4.2) for each junction in which the expression for the currents from Eq. (4.3) are used. For generality the sums are considered to extend from 1 to $(J - 1)$ to take into account any possible branches (simple parallel branches are considered as one)

$$0 = \sum_{i \neq l} i_{il} = \sum_{i \neq l} [(V_l - V_i)G_{il} + \mathcal{E}_{il}G_{il}]$$
$$= V_l \sum_{i \neq l} G_{il} - \sum_{i \neq l} G_{il}V_i + \sum_{i \neq l} G_{il}\mathcal{E}_{il}$$

[1] Neglecting simple parallel branches, the most complicated circuit of J junctions involves a branch between each pair of junctions, and the maximum number of branches is thus

$$B_c = \frac{J(J - 1)}{2}$$

The simplest network of J junctions involves no more than three branches to each junction, or by induction

$$B_s = 3(J - 1) - 3$$

$J \geq 3$ or the cases are trivial. From these equations it is evident that the number of mesh equations may be as great as $\frac{1}{2}(J - 1)(J - 2)$ or as small as $2(J - 1) - 3$. In the first case the number of series equations exceeds the number of shunt equations by $\frac{1}{2}(J - 1)(J - 4)$ and in the second by $(J - 4)$.

where G_{il} has been written for $1/R'_{il}$. Using the conventions

$$I_l = \sum_{i \neq l} G_{il}\mathcal{E}_{il} \qquad \text{and} \qquad G_{ll} = -\sum_{i \neq l} G_{il}$$

the equations for each of the $(J - 1)$ junctions become

$$I_l = \sum_{i=1}^{i=J-1} G_{il}V_i \tag{4.4}$$

If the network contains parallel branches, the G_{il} between a pair of junctions is clearly the sum of the conductances of the parallel branches between the junctions. Given the values of the conductances of all branches of a network and the values of any emfs. they contain, Eqs. (4.4) can be written down by inspection. If the currents leaving one junction or the potential difference between it and neighboring junctions are of particular interest, the potential of that junction is chosen as zero. The determinantal method of solving these $(J - 1)$ equations is by far the simplest way of determining the potentials of the junctions referred to that chosen as zero in terms of the G's and I's.[1] Writing D' for the determinant formed by the array of G_{il}'s and D'_i for the altered determinant in which the column of coefficients of V_i has been replaced by the column of I_l's, then $V_i = D'_i/D'$. The physical nature of the problem insures that D' does not vanish, and it should be noted that $G_{il} = G_{li}$. Expanding D'_i in terms of the I_l's and their cofactors, say B_{il}, which are the same as the cofactors of G_{li} in D,

$$V_i = \sum_{l=1}^{l=J-1} \frac{B_{il}}{D'} I_l \tag{4.5}$$

The problem is then solved completely for the network, although the algebraic work is tedious for many junctions. In encountering Eqs. (4.4) and (4.5) for the first time it is well to write them out completely, putting in numbers for the letters, choosing a simple case of four or five junctions, and not forgetting the appropriate signs for the cofactors. An illustrative example is given after the discussion of series analysis.

Series analysis proceeds by writing down equations of the form of Eq. (4.1) for each mesh in which the current through each resistance is the algebraic sum of the postulated mesh currents through that resistance. The M meshes may be chosen quite arbitrarily as long as each branch is traversed at least once. If currents in certain branches are of special interest, the meshes should be chosen as far as possible so that these branches are traversed but once. The mesh currents chosen are then the currents in these branches. The emfs. encountered in traversing the

[1] See any algebraic text on methods of manipulating determinants.

mesh, the algebraic sums of which appear on the left of the equations, are reckoned positive if they tend to send current in the assumed direction of the mesh current. The sense of the latter is arbitrary, but it is generally chosen clockwise; and if on solution it is found to be negative, the current actually flows in a counterclockwise sense. Let $\mathcal{E}_k$ be the net emf. encountered in traversing the kth mesh and R_{kk} the total resistance through which the postulated current i_k in this mesh flows. The quantity R_{jk} is written for the resistance common to mesh j and k through which both mesh currents i_j and i_k flow. Then the M network equations are

$$\mathcal{E}_k = \sum_{j=1}^{j=M} R_{jk} i_j \qquad (4.4')$$

FIG. 4.4.—Analysis of a three-mesh network.

As in the case of the shunt discussion the solution for the i_j's is readily written down formally in terms of the determinant D of the array of R_{jk}'s and the A_{jk}'s that are the cofactors in this determinant of the R_{kj}'s.

$$i_j = \sum_{k=1}^{k=M} \frac{A_{jk}}{D} \mathcal{E}_k \qquad (4.5')$$

Thus the results of the shunt and series analyses are formally very similar. The algebraic work may be tedious but is very straightforward, and the formal expressions of Eqs. (4.4′) and (4.5′) are very useful in deriving certain general circuit theorems.

As an example consider the circuit of Fig. 4.4. It is composed of six branches and has four junctions. Thus there would be the same number of equations, namely, three, in either a shunt or a series analysis. First of all, however, the equations may be written down in terms of branch currents and branch resistances:

$$\begin{aligned}
\mathcal{E}_d &= i_d R_d - i_f R_f + i_b R_b \\
\mathcal{E}_d &= i_d R_d - i_e R_e - i_c R_c \\
\mathcal{E}_a &= i_f R_f + i_a R_a - i_e R_e
\end{aligned}$$

And at the junctions

$$\begin{aligned}
-i_a + i_b + i_f &= 0 \qquad & i_d - i_b + i_c &= 0 \\
i_e - i_c + i_a &= 0 \qquad & i_d + i_e + i_f &= 0
\end{aligned}$$

Eliminating i_d, i_e, and i_f yields

$$\begin{aligned}
\mathcal{E}_a &= (R_a + R_e + R_f) i_a - R_f i_b - R_e i_c \\
\mathcal{E}_d &= -R_f i_a + (R_b + R_d + R_f) i_b - R_d i_c \\
-\mathcal{E}_d &= -R_e i_a + R_d i_b + (R_d + R_c + R_e) i_c
\end{aligned}$$

These three equations may be solved for i_a, i_b, and i_c by ordinary methods, and the central equations may be used to determine the other three currents. In terms of the shunt and series analyses, respectively, the equations are seen to be

$$\begin{aligned}\mathcal{E}_d G_d &= -(G_b + G_e + G_d)V_1 + G_b V_2 + G_c V_3 \\ -\mathcal{E}_a G_a &= G_b V_1 - (G_a + G_b + G_f)V_2 + G_a V_3 \\ \mathcal{E}_a G_a &= G_c V_1 + G_a V_2 - (G_a + G_c + G_e)V_3\end{aligned}$$

where the G's are written for the reciprocals of the R's and

$$\begin{aligned}\mathcal{E}_a &= (R_a + R_f + R_e)i_1 - R_f i_2 - R_e i_3 \\ \mathcal{E}_d &= -R_f i_1 + (R_b + R_d + R_f)i_2 - R_d i_3 \\ -\mathcal{E}_d &= -R_e i_1 - R_d i_2 + (R_c + R_d + R_e)i_3\end{aligned}$$

The symmetry of the determinants D' and D, which results from the conditions $G_{il} = G_{li}$ and $R_{jk} = R_{kj}$, is evident and is of considerable assistance in verifying the correctness of the analyses.

It is clear from Eq. (4.5′) that each emf. acts separately in producing a component current in each branch of the network. In fact i_j can be written as the sum of elementary currents i_{jk}, where $i_{jk} = \kappa_{jk}\mathcal{E}_k$. κ_{jk} which is written for the quotient A_{jk}/D is known as the *transfer conductance*. Here i_{jk} is the current in mesh j produced by the emf. associated with mesh k. This consequence of the linearity of Eq. (4.4′) or (4.5′) is called the *superposition theorem*. It may be stated in words by saying that each emf. in a network acts independently of all the rest in producing the network currents. If i_a is the current flowing in branch a and an additional emf. $\mathcal{E}_b$ is inserted in branch b, the current then flowing through branch a is i_a plus the current emf. $\mathcal{E}_b$ would send through branch a in the absence of any other emfs.

The simplest case is, of course, that in which but one emf. is present in the network. For example, in Fig. 4.5 the emf. $\mathcal{E}_1$ in branch 1 produces a current i' in branch 6. If the emf. $\mathcal{E}_2$ in branch 2 produces a current i'' in branch 6, the current that would flow there with both emfs. present would be $i' + i''$. The sources of emf. have, of course, certain resistances associated with them and these resistances must not be neglected in setting up the circuit equations. When only one emf. is present in a network, say the emf. $\mathcal{E}_1$ in mesh 1, the current flowing in any branch j is given by

$$i_j = \frac{A_{j1}}{D}\mathcal{E}_1$$

The ratio $\mathcal{E}_1/i_1$ is the effective resistance presented by the network to the emf. Thus D/A_{11} is the input resistance of the network as viewed from the terminals of an emf. present in branch 1. A network which contains no sources of emf. is known as a *passive network*. And from

this discussion it is evident that the resistance presented by a passive network to an emf. applied between any two points is obtained by setting up the equations for the resulting network and solving for D/A_{11} where the branch containing the emf. is reckoned as branch 1.

Another useful theorem is a consequence of the fact that the resistance common to meshes j and k is the same as the resistance common to meshes k and j, *i.e.*, $R_{jk} = R_{kj}$. Thus the determinant D is symmetrical about the diagonal from the upper left- to the lower right-hand corner and A_{jk} is seen to be equal to A_{kj}. Therefore the transfer conductance from mesh j to mesh k is the same as that from mesh k to mesh j and an emf. placed in branch 1 produces the same current in branch 2 as it would produce in branch 1 if it had been placed in branch 2. As an

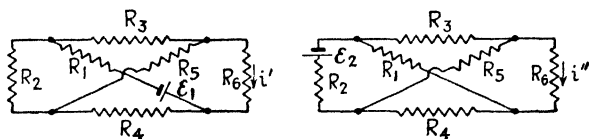



FIG. 4.5.—Superposition and reciprocity theorems. If both batteries are present the current through R_6, say, is $i = i' + i''$.

example, on referring to Fig. 4.5, it can be stated from inspection that if i is the current in branch 2 in the left-hand circuit, the current in branch 1 of the right-hand circuit is equal to $i\mathcal{E}_2/\mathcal{E}_1$. This implication of the equality $R_{jk} = R_{kj}$ is known as the *reciprocity theorem.* It introduces many simplifications in the analysis of complicated circuits. It may be stated in terms of resistanceless generators and ammeters instead of in terms of emfs. and currents. The generator and ammeter, however, need not be resistanceless for the theorem to apply. For it is evident that if they have the same finite resistance, the determinant D will be unaltered if their positions are interchanged and such an interchange will not affect the reading of the ammeter.

Assume that the equations for a given network have been solved and the currents in each of the branches determined. Now consider the result of inserting a resistance δR_1 in one of the branches, say branch a which appears only in mesh 1. The equations for the network are then altered to the extent that the coefficient of i_1 in the equation for mesh 1 becomes $R_{11} + \delta R_1$. If this additional term is taken to the left of the equality sign, it is seen to be equivalent to an additional emf. $-\delta R_1 i_1$, *i.e.*, if an additional emf. $\delta\mathcal{E} = \delta R_1 i_1$ were also inserted in the branch, the currents would everywhere be the same as before the alteration. Thus, by the superposition theorem the effect of the change δR_1 is the same as if an emf. $-\delta R_1 i_1$ had been inserted in the branch in series with δR_1. This result is known as the *compensation theorem* and may be stated in general by saying that if any change is made in the resistance of a

branch, the effect on all the mesh currents is the same as if an emf. equal to minus the product of the change in resistance and the branch current had also been inserted. This is a very useful theorem for complicated circuits. Referring to Fig. 4.6, the insertion of the resistance r_6 is equivalent to the insertion of an emf. of $-r_6 i_6$ in series with r_6 where i_6 is the current initially flowing in that branch.

The potential difference that appears across a resistance in a network, let us say R_j, which appears only in the jth mesh, is

$$V_j = R_j i_j = R_j \sum_{k=1}^{k=M} \frac{A_{jk}}{D} \mathcal{E}_k$$

by Eq. (4.5′). If R_j becomes indefinitely large, which corresponds to opening the network at this point, the other resistances which appear

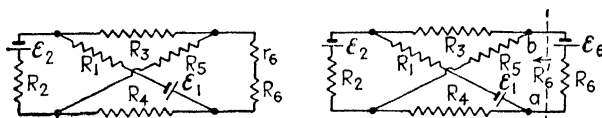



FIG. 4.6.—Compensation and Thévenin's theorems.

in D summed with R_j can be neglected in comparison with this resistance and D approaches the value $R_j A_{jj}$. Thus V_j', which is the open-circuit potential difference appearing across the break in the network at this point, is given by canceling R_j in numerator and denominator as

$$V_j' = \sum_{k=1}^{k=M} \frac{A_{jk}}{A_{jj}} \mathcal{E}_k \tag{4.6}$$

But recalling that D/A_{jj} is the input resistance, R_j', of the network as viewed from the several terminals, on comparing Eqs. (4.5′) and (4.6) it is seen that

$$i_j = \frac{V_j'}{R_j'}$$

This means that the current which will flow through a resistanceless connection made between any two points of a network is equal to the open-circuit potential difference appearing between these points divided by the input resistance of the network as viewed from these points. Thus, if two available terminals emerge from a network, they may be considered, as far as external circuit calculations are concerned, as the terminals of a simple series circuit of an emf. equal to the potential difference appearing across them and a resistance equal to the input resistance as measured between them. This is known as *Thévenin's*

theorem and is one of the most useful theorems for the analysis of complicated networks. As an example consider the right-hand portion of Fig. 4.6. Assume first that $\mathcal{E}_6$ and R_6 are removed and let V_6' be the potential difference between the terminals a and b and R_6' the resistance of the network as measured between these terminals. If the branch containing $\mathcal{E}_6$ and R_6 is added, the current that will flow through it is the algebraic sum of V_6' and $\mathcal{E}_6$ divided by the sum of R_6' and R_6.

Power-transfer Theorem.—Consider any two-terminal network which is equivalent to an emf. $\mathcal{E}$ and a resistance R_e connected to a passive two-terminal network of input resistance R_l. The circuit is represented schematically in Fig. 4.7. The current flowing between the networks is $\mathcal{E}/(R_e + R_l)$ and the power supplied to the load R_l is equal to iV_l or i^2R_l. Writing this as P_l

$$P_l = \frac{\mathcal{E}^2 R_l}{(R_e + R_l)^2} \tag{4.7}$$

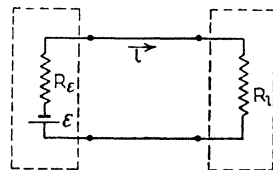



FIG. 4.7.—Fundamental power-transfer theorem.

Thus the power delivered to the load is proportional to $\mathcal{E}^2$. If R_e is variable, the power supplied increases as R_e decreases, being a maximum for $R_e = 0$. The power delivered to the load may also be considered as a function of the load resistance and it is seen to be vanishingly small for either very large or very small values of R_l. The value of R_l which makes P_l a maximum is found by setting the derivative of P_l with respect to R_l equal to zero. This yields

$$\frac{dP_l}{dR_l} = 0 = \mathcal{E}^2\left(\frac{1}{(R_e + R_l)^2} - \frac{2R_l}{(R_e + R_l)^3}\right)$$

or

$$R_l = R_e \tag{4.8}$$

The fact that this condition represents a maximum power transfer may be verified by forming the second derivative of P_l with respect to R_l and observing that it is negative when Eq. (4.8) holds. Equation (4.8) is a very important one and may be stated in words by saying that a given power source can deliver most power to a load if the load resistance is equal to the internal resistance of the power source. When Eq. (4.8) holds, the resistance of the load and power source are said to be *matched.*

On inserting the value of R_l from Eq. (4.8) into Eq. (4.7) the maximum power that can be delivered by the source is seen to be

$$P_{ml} = \frac{\mathcal{E}^2}{4R_e}$$

The total power developed in the circuit is $\mathcal{E}^2/(R_e + R_l)$, which is $\mathcal{E}^2/2R_e$ when Eq. (4.8) is satisfied. Thus when resistances are matched, half

the power developed is delivered to the load and half is lost in the source. The fraction of the total power developed that is delivered to the load increases as the ratio R_e/R_l decreases. The ratio of the power delivered to a load to the maximum power that could be delivered by the source is

$$\frac{P_l}{P_{ml}} = \frac{4R_lR_e}{(R_e + R_l)^2} \tag{4.9}$$

This ratio does not vary rapidly with R_l in the neighborhood of R_e; in fact, if α is the percentage difference between R_l and R_e $\left(i.e., \alpha = \frac{(R_e - R_l)}{R_l}\right)$, the fractional difference in power transfer is only $\alpha^2/4$ if α is small. Thus the percentage difference in power transfer becomes small as the square of α. This implies that very accurate resistance matching is frequently an unnecessary refinement.

Power ratios and also voltage and current ratios are often measured most conveniently in terms of their logarithms. The reasons for this will appear in connection with cascade arrangements of four-terminal networks. Since the ratio [Eq. (4.9)] is less than one, the logarithm of its reciprocal is used with a negative sign. This logarithm is said to be the *transfer power loss* between the two networks. When the resistances are matched, the transfer power loss is zero. When the logarithm is taken to the base e, the power loss is said to be in *nepers*. When the logarithm is taken to the base 10, the loss is said to be in *bells*. The commonest unit is a tenth of a bell or a decibel, commonly written *db*. Thus the transfer power loss in decibels from Eq. (4.9) is

$$-10 \log_{10}\left(\frac{P_{ml}}{P_l}\right) = 20 \log_{10} (R_e + R_l) - 10 \log_{10} (4R_eR_l)$$

4.4. The Passive Four-terminal Network (Passive Quadripole).—The problem of the transfer of power from the two terminals of a power source to the two terminals of a load is a very important one. In order to effect such a transfer a device with two pairs of terminals must be used. The transfer circuit itself generally contains no source of power and hence is known as a *passive quadripole*. In Fig. 4.7 it is represented merely by the two straight lines joining the source and load. In the simple case of two short connections the only effect produced is the introduction of a resistance equal to that of the two connections. However, when the power must be transferred over a great distance, the resistance of the conductors becomes appreciable and also the shunt paths introduced by supporting insulators may be important. Long power-transfer circuits of this type are known as *lines* or *cables*. A second type of transfer network is often inserted between two circuits in order to reduce the power transfer by a given amount. It is frequently desirable to do this without changing the effective resistance presented to either source or load. This is particularly true if either of these contain nonlinear elements which do not obey Ohm's law. The situation is most frequently encountered in communication circuits, and the device

for decreasing power transfer without affecting the terminal resistances is known as an *attenuator*.

The Attenuator.—It can be shown that as far as the external circuits are concerned, any passive quadripole can be represented by a simple network of three resistances. One method of arranging these resistances is shown in the central portion of Fig. 4.8. From the form of this network it is known as a *T section* and the arms X_1 and X_2 and the stem Y are in general all unequal resistances. It is evident that the network as observed from either set of terminals is unaltered if the series arms X_1 and X_2 were each divided in half, one-half being placed in series with the lower of each pair of terminals. From its form such a network is known as an *H section*. To external observation the two types are identical, but the latter possesses a higher degree of symmetry within the network and is employed when such symmetry is of importance. The T-section attenuator is actually constructed in the schematic form of Fig. 4.8. For simplicity this discussion will be limited to the symmetrical case in which $R_e = R_l$ and the input resistances from either pair of terminals, R_a and R_b, are equal to one

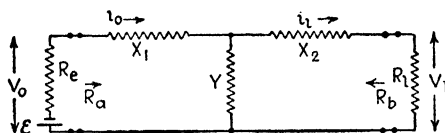



FIG. 4.8.—T-section attenuator.

another and to the load or source resistance. For these conditions to be fulfilled it is evident that X_1 must also equal X_2. Writing X for the series resistances, R for the terminal resistances, and Y for the shunt resistance, the condition for the input resistance to equal the terminal resistance is

$$R = X + \frac{Y(X + R)}{Y + X + R}$$

Or converting X and Y into units of R, *i.e.*, $x = X/R$ and $y = Y/R$, this condition can be written

$$2xy = 1 - x^2 \tag{4.10}$$

The quantities x and y are also related through the ratios i_0/i_l or V_0/V_l. From Fig. 4.8

$$\frac{i_l}{i_0} = \frac{V_l}{V_0} = \frac{y}{x + y + 1} \tag{4.11}$$

Eliminating y between Eqs. (4.10) and (4.11)

$$\frac{V_0}{V_l} = \frac{i_0}{i_l} = \frac{1 + x}{1 - x} \tag{4.12}$$

Ten times the $\log_{10}$ of the ratio of the input power to the output power is said to be the attenuation, α, in decibels of the quadripole. Thus

$$\alpha = 10 \log_{10} \frac{V_0 i_0}{V_l i_l}$$

or

$$\frac{V_0 i_0}{V_l i_l} = 10^{\frac{\alpha}{10}} = e^{\frac{c\alpha}{10}}$$

where $c = 2.3026$ which is $\log_e 10$, where e is the base of natural logarithms. Writing s for $c\alpha/40 = 0.05756\alpha$ and solving Eq. (4.12) for x

$$x = \frac{e^{2s} - 1}{e^{2s} + 1} = \frac{e^s - e^{-s}}{e^s + e^{-s}} = \tanh s \tag{4.13}$$

and from Eq. (4.10)

$$y = \frac{1 - x^2}{2x} = \frac{8}{e^{2s} - e^{-2s}} = \text{csch } 2s \tag{4.14}$$

Thus the proper values of x and y for any desired attenuation α can be obtained directly from tables of the hyperbolic functions. These values of x and y maintain the

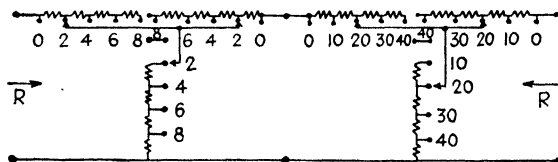



FIG. 4.9.—Two T-section attenuators in cascade for obtaining up to 48 db attenuation in steps of 2 db.

desired constant input and output resistances and produce an attenuation α. The voltage or current ratio is e^{2s} which is the square root of the power ratio, e^{4s}.

A variable attenuator can be constructed by the use of a three-arm multipole switch as indicated by either one of the units in Fig. 4.9. One arm moves over contacts on the y branch and the other two over the x branches. The contacts are generally so arranged that the attenuation occurs in equal steps. Any number of attenuators can be connected to follow one another in cascade, provided they are all designed to operate between the terminating resistances. Since the ratio of the power output to power input for the cascade is equal to the product of this ratio for the separate units, the attenuation which is the logarithm of this ratio is equal to the sum of the attenuations for each unit separately. Figure 4.9 represents schematically two attenuators of five steps each which together can produce a variable attenuation in steps of 2 db from 0 to 48 db. Thus, if the potential difference appearing across the input terminals is 1 volt and 1 amp. is the input current, the output power can be varied from 1 watt to $10^{-4.8}$ or about 16 μw. The potential appearing across the load and the current through it can be varied from the input values to about 4 mv. or 4 ma., respectively.

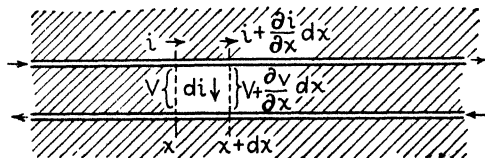



FIG. 4.10.—Uniform cable with leakage.

Lines and Cables.—In a two-conductor cable the copper wires are generally separated by rubber or paper insulation and the leakage resistance per unit length between them is continuous and uniform along the cable. In the case of lines supported by poles the high-resistance leakage paths are present only at the supporting insulators, but these generally occur at regular intervals along the line. If the line is long in comparison with the distance between supports, the circuit analysis is very similar to that for a continuous cable. Figure 4.10 represents schematically a section of a two-

conductor cable. V is the potential difference between the conductors at a distance x from the input of the cable and i is the current flowing through either conductor at that point. Assuming a continuous variation of V and i, the potential difference and current a distance dx farther along the line are given by Taylor's theorem as $V + (dV/dx)dx$ and $i + (di/dx)dx$. The difference between the potential difference at x and at $x + dx$ is due to the iR drop between these two points, which is equal to $iR'\,dx$ if R' is the resistance per unit length of the conductor pair (twice the resistance per unit length of a single wire). Similarly the difference between the conductor current at x and $x + dx$ is the leakage current through the insulator between these two points, which is equal to $VG'\,dx$ if G' is the conductance or reciprocal of the leakage resistance per unit length of the insulating material.

Stating these results in the form of equations

$$i - \left(i + \frac{di}{dx}dx\right) = G'V\,dx \qquad \text{or} \qquad \frac{di}{dx} = -G'V \tag{4.15}$$

$$V - \left(V + \frac{dV}{dx}dx\right) = R'i\,dx \qquad \text{or} \qquad \frac{dV}{dx} = -R'i \tag{4.16}$$

On differentiating one or the other with respect to x and eliminating i or V both these quantities are found to obey a differential equation of the form

$$\frac{d^2V}{dx^2} = R'G'V \tag{4.17}$$

The solution of this equation can be written most conveniently in terms of the hyperbolic functions as

$$V = A_1 \sinh \alpha x + A_2 \cosh \alpha x \tag{4.18}$$

where α is an abbreviation for $\sqrt{R'G'}$. And from Eq. (4.16)

$$i = -g(A_1 \cosh \alpha x + A_2 \sinh \alpha x) \tag{4.19}$$

where g is an abbreviation for $\sqrt{G'/R'}$ which is a quantity of the dimensions of a conductance. Let $i = i_0$ and $V = V_0$ at $x = 0$, the input end of the cable. Then from Eqs. (4.18) and (4.19), $i_0 = -gA_1$ and $V_0 = A_2$, or

$$V = -\frac{i_0}{g} \sinh \alpha x + V_0 \cosh \alpha x$$

$$i = i_0 \cosh \alpha x - gV_0 \sinh \alpha x \tag{4.20}$$

These equations give the potential difference and current at any point on the line in terms of the input potential and current.

For many purposes it is more convenient to have these quantities in terms of the terminating load resistance. If this has the value R_l and is located at a distance $x = l$ from the input of the cable, V must equal iR_l when $x = l$. Inserting l for x in Eqs. (4.20) and equating the resultant V to $R_l i$ yields the following ratio of V_0/i_0:

$$\frac{V_0}{i_0} = \frac{1}{g} \tanh (\alpha l + \phi) \tag{4.21}$$

where ϕ is the angle whose hyperbolic tangent is $R_l g$ ($\tanh \phi = R_l g$). Eq. (4.21) gives the input resistance of the line in terms of the parameters α and ϕ and the terminating resistance R_l. Using this equation to eliminate i_0, Eqs. (4.20) can be written

$$V = V_0 \frac{\sinh [\alpha(l - x) + \phi]}{\sinh (\alpha l + \phi)}$$

$$i = gV_0 \frac{\cosh [\alpha(l - x) + \phi]}{\sinh (\alpha l + \phi)} \tag{4.22}$$

These are the most convenient equations for determining the line potential and current at any point if R', G', and R_l are known. The power put into the line is $i_0 V_0$ which from Eq. (4.21) is given by

$$V_0 i_0 = V_0^2 g \coth (\alpha l + \phi)$$

and the power delivered to the load is iV at $x = l$ or, from Eqs. (4.22)

$$V_l i_l = V_0^2 \frac{g}{2} \frac{\sinh 2\phi}{\sinh^2 (\alpha l + \phi)}$$

The ratio of these two gives the fraction of the input power delivered to the load which is $(\sinh 2\phi)/[\sinh 2(\alpha l + \phi)]$. These equations are all very useful in the analysis of direct-current cables and they may also be used for lines if l is great in comparison with insulator spacings. In this case if R is the line resistance between insulators and r the leakage resistance at an insulator, $\alpha = (1/d)\sqrt{R/r}$, $g = 1/\sqrt{Rr}$, and if m is the total number of line sections between insulators and V and i refer to the current at the nth section, $l = md$ and $x = nd$, where d is the length of a section between insulators.

4.5. Laboratory Resistances.—Though currents and potential differences and their product, which is the power, are the quantities which are of primary interest, the wide applicability of Ohm's law makes resistance of great importance also. If R is known, the equations $V = Ri$ and $P = iV$ may be used to determine any two of the quantities V, i, and P in terms of the third. Also it is easier to prepare, maintain, and compare standards of resistance than for any of the other electrical quantities. It is evidently necessary to have a practical standard of one of the other three quantities as well and of these it is most convenient to choose V. The absolute calibration of the practical standards of R and V in terms of the ultimate standards of length, mass, and time is deferred till Secs. 9.5 and 10.4. The methods of comparing resistances will form the subject of the following section on bridge circuits, and methods of comparing potentials will be taken up in the subsequent section. Here a few of the typical forms of laboratory resistance will be considered.

Constant Resistances.—A laboratory is generally equipped with primary precision resistance standards in terms of which other resistances can be measured. The requirements for such a standard are that it shall be permanent and definite and show only small variations for changes in the conditions under which it is used. A great deal of research has been carried on in standardizing laboratories for the development of satisfactory standards and these are now available in values from 0.0001 to 10,000 ohms.[1] Permanence requires that the standards be very carefully constructed of high-grade materials and that they be main-

[1] For a complete discussion of electrical standards the reader is referred to Curtis, "Electrical Measurement," McGraw-Hill Book Company, Inc., New York, 1937.

tained under constant conditions. The latter can be partially accomplished by the use of protective coatings, but it is most satisfactory to seal hermetically the standards in a metal container. The problem of definiteness is of importance only for resistances of less than 10 ohms. The distribution of current over a cross section of the resistance element is not entirely independent of the method of making the terminal contacts, but this becomes of importance only for very low resistances. The major cause of resistance variation is temperature and as a consequence standard resistances are used in a constant-temperature bath. Also the resistance elements are composed of an alloy such as manganin or constantan which has a zero temperature coefficient in the ordinary temperature range. When in use, heat is of course developed in the resistor and the design must be such as to afford good heat interchange with the surroundings. Also, if the junctions between the resistance element and the copper terminals are not at exactly the same temperature, a net thermal emf. will appear across the resistance (see Sec. 6.4). This source of error can be minimized by choosing an alloy such as manganin for which the thermoelectric power relative to copper is only a few microvolts per degree centigrade. All thermal sources of error are reduced by keeping the power dissipated in the resistance as low as possible.

FIG. 4.11.—Section through a resistance standard.

Figure 4.11 illustrates schematically a section through a typical standard resistance. The terminals are heavy copper rods with their ends bent over in such a way that they may be set in mercury cups which form part of the external circuit. The ends of the rods are amalgamated to insure good contact and generally additional binding posts are provided for so-called potential terminals. The ratio of the potential difference between these terminals to the current flowing through the element is defined as the resistance of the unit. The copper terminals pass through an insulating disk to which they are rigidly attached. To this disk is also attached a supporting cylinder upon which the resistance winding is wound. It also serves as the lid of a metal protective container for the element. This container may be filled with paraffin oil and hermetically sealed or it may have holes through the sides for the circulation of oil at a constant temperature. Provision is also made for the introduction of a thermometer. The resistance element itself is generally a manganin wire of the proper length and cross section which is wound on the supporting cylinder and hard-soldered to the copper-rod terminals.

Standard resistances of this general type may be relied upon to have a resistance which is constant over long periods of time to within a few parts in 10^5.

Ammeter shunts and series resistances for voltmeters are of the nature of semistandard resistances. They are also made of manganin or some other material with a low temperature coefficient of resistance, and when they have to dissipate more than about 1 watt an arrangement for the circulation of air is generally provided. Other fixed resistances are generally constructed by winding manganin wire on brass or wooden spools which are then coated with shellac or paraffin. In order that these may be used for alternating as well as direct currents they are generally wound noninductively (see Sec. 13.3). Less reliable resistors can be constructed using low-conductivity ceramic materials. Mixtures of graphite and clay extruded in the form of rods and subsequently baked are used for the construction of resistors in the range from 1 to 10^7 ohms. These are inexpensive and convenient, but they have a large temperature coefficient and are less permanent than wire-wound resistors. Very high resistances in the range from 10^7 to 10^{13} ohms can be made by drawing an India-ink line on good-quality paper, though these are generally unreliable even for short periods. More satisfactory resistances in this range can be constructed by sputtering or evaporating thin metallic layers on long quartz or glass rods and placing them in an evacuated container.

FIG. 4.12.—Plug type of resistance box.

Variable Resistances.—Variable resistances are generally made by suitable combinations of fixed resistances so connected that the effective resistance presented by the two terminals of the combination can be varied in appropriate steps. The construction of the individual fixed resistances have been described above and the arrangement for varying the effective resistance generally takes one of two forms. In the plug type of resistance box the resistance units are usually connected in series and the junctions between units are brought to heavy brass blocks on the top of the box. A series of plugs can be inserted between these blocks, short-circuiting the units. Such an arrangement is shown schematically in Fig. 4.12. If the resistances are in order 1, 1, 3, 5, 10, 20, and 30 ohms, by a suitable arrangement of plugs the resistance between the terminals can be varied in 1-ohm steps from 0 to 70 ohms. Other plug-box circuits can be designed in which one plug is sufficient for each decade variation. The units chosen are seldom smaller than 0.1 ohm, for otherwise the uncertainty involved in the resistances of the plug contacts is an appreciable fraction of the lowest unit. The plugs must always be tightly

inserted to minimize this source of error. Such boxes are generally accurate to 0.1 per cent. They are designed for precision rather than current-carrying capacity and in use a box should not be called upon to dissipate more than about 1 watt.

Another arrangement for varying the number of resistance units in series with the terminals of a resistance box is the use of single-arm

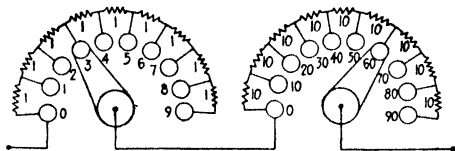



FIG. 4.13.—Schematic representation of a two-decade dial resistance box.

multipole switches of the dial type. There are generally 10 poles per switch, each switch covering a decade of resistance. Two decades are shown in Fig. 4.13, covering a range from 0 to 99 ohms in 1-ohm steps. If the resistance steps are as small as 0.1 ohm, the switches must be carefully made to insure low contact resistance. The upper limit of such a box is generally about 100,000 ohms and the permissible power dissipation is about the same as for plug boxes. Shunts for varying the sensitivity of a galvanometer are frequently made in the dial-switch form also, but the resistance steps are unequal. The Ayrton type of galvanometer shunt is illustrated schematically in Fig. 4.14. Let R be the total value of the series of resistances in the shunt and R_g the resistance of the galvanometer. Then, if i is the current flowing in the line to the shunt and i_g is that which flows through the galvanometer, by Kirchhoff's laws

$$(i - i_g)nR = i_g[R_g + (1 - n)R]$$

or

$$i_g = \frac{n}{1 + \frac{R_g}{R}} i$$

FIG. 4.14.—Schematic representation of an Ayrton shunt.

where n is the fraction of R directly in shunt with the line. If R_g/R is negligible in comparison with unity, n is the ratio i_g/i. The value of n is marked at each switch pole and since the deflection of the galvanometer is proportional to i_g, this number represents the fraction of the full sensitivity of the galvanometer corresponding to that switch setting. Thus in Fig. 4.14 the available fractional sensitivities are: 0, 0.01, 0.1, and 1, there being one switch setting that open-circuits the line. This is the most convenient device for varying a galvanometer's sensitivity. In addition it keeps the resistance across the galvanometer terminals approximately

constant, and if this is chosen to be the critical damping resistance, the use of the instrument is greatly facilitated (see Sec. 10.5). The ratio R_g/R is generally of the order of 0.01, hence the ratio i_g/i is equal to n to within about 1 per cent. If a greater accuracy is required, the resistance ratio must be known and taken into account.

There are many types of variable resistance which are designed for the dissipation of power rather than for accuracy. In one form oxidized resistance wire such as nichrome is wound on an insulating cylinder and a sliding contact is arranged to traverse its length. The oxide provides insulation between turns and is scraped off the outer surface of the wire in the path of the sliding contact. Large resistances of this type will dissipate several hundred watts, but their resistance changes considerably with the temperature. Variable low resistances are frequently made by placing a series of carbon blocks in the circuit and varying the force with which they are pressed together. This varies the contact resistance between the blocks and hence the total resistance presented by the pile. High resistances can be made on the same general principle or by preparing a thinly carbonized high-resistance layer on the surface of an insulator and using a sliding contact to vary the portion included in the circuit. However, high resistances of this type are generally unreliable and will at best dissipate only a few watts.

FIG. 4.15.—Wheatstone bridge.

4.6. Bridge Circuits.—The basic circuit for the comparison of resistances is the *Wheatstone bridge,* which is illustrated schematically in Fig. 4.15. From the diagram the equations of the bridge are seen to be the following:

$$\begin{aligned} \mathcal{E} &= (R_3 + R_4 + R_e)i + R_3 i_g - (R_3 + R_4)i' \\ 0 &= R_3 i + (R_1 + R_3 + R_g)i_g - (R_1 + R_3)i' \\ 0 &= -(R_3 + R_4)i - (R_1 + R_3)i_g + (R_1 + R_2 + R_3 + R_4)i' \end{aligned} \tag{4.23}$$

On solving for i_g it is seen to be equal to $\frac{\mathcal{E}(R_1R_4 - R_2R_3)}{D}$, where D is the determinant of the coefficients of the currents. Thus the condition that i_g shall be equal to zero, *i.e.*, that there shall be no galvanometer deflection, is evidently

$$R_2R_3 = R_1R_4 \qquad \text{or} \qquad R_2 = \frac{R_1R_4}{R_3} \tag{4.24}$$

This is known as the balance condition of the bridge and as it is determined by a null deflection of the galvanometer, it does not involve a

knowledge of either emfs. or currents. Thus if the balance condition is established and, say, R_1 and the ratio R_4/R_3 are known, the value of R_2 is determined. In one type of bridge, R_1 has a fixed known value and the ratio R_3/R_4 is variable continuously by moving a sliding contact along a resistance wire, the ratio being given by the lengths of the wire in the two arms. This is known as a *slide-wire bridge*. In another type the ratio is variable in large steps and the balance is achieved by varying R_1. In either type a tapping key is included in the galvanometer circuit to test for the balance condition at various settings.

From the reciprocity theorem it is evident that the relative positions of $\mathcal{E}$ and G in the circuit could be interchanged and the balance condition would still be given by Eq. (4.24). However, the sensitivity of the bridge, which may be taken as the galvanometer deflection for a small degree of unbalance, may be quite different with $\mathcal{E}$ and G interchanged. The deflection of a galvanometer is proportional to the current flowing through it and the constant of proportionality is itself proportional to the square root of the galvanometer resistance (Sec. 10.5). Thus, writing d for the deflection, $d = C\sqrt{R_g}\, i_g$, where C is approximately a constant for a given galvanometer design. From this it is evident that the square of the deflection is proportional to the joule heating in the galvanometer. The deflection will be a maximum when the power dissipated in the instrument is a maximum and the general power theorem shows that the deflection will be a maximum subject to a constant emf. and bridge resistance when the resistance of the galvanometer is matched to that of the output of the bridge. In Fig. 4.15 the resistance presented by the balanced bridge to the galvanometer terminals is the resistance in parallel of the two series branches $(R_1 + R_3)$ and $(R_2 + R_4)$, for owing to the balance condition no current would flow through R_e owing to an emf. in circuit g. This is a satisfactory approximation even when the bridge is slightly unbalanced and hence the output resistance of the bridge is $\dfrac{(R_1 + R_3)R_4}{(R_3 + R_4)}$. If the positions of $\mathcal{E}$ and G were reversed, the output resistance would be $\dfrac{(R_3 + R_4)R_1}{R_1 + R_3}$. These output resistances are in general quite different and it is advantageous to place the galvanometer in the position for which the match with R_g is best. If various galvanometers are available, the one should be chosen which is best suited to the bridge in one position or the other. Of course, the sensitivity can also be increased by decreasing the magnitude of the bridge resistances or by increasing $\mathcal{E}$. But generally the first of these is infeasible and as both lead to an increased power expenditure in the bridge, they cannot be carried beyond a certain point.

The problem of optimum bridge design subject to a maximum safe power dissipation in the bridge and assuming a perfectly matched galvanometer may be solved in the following way. Consider that R_2 differs slightly from the proper value for balance, *i.e.*, $R_2 = \frac{(1+\alpha)R_1R_4}{R_3}$ where α is so small that its square may be neglected in comparison with unity. Inserting this value in Eqs. (4.23) and solving for i_g in terms of i

$$i_g = \frac{R_1R_4\alpha i}{2(R_1 + R_3)(R_2 + R_4)}$$

where the output resistance of the bridge has been inserted for R_g. Eliminating i by means of the input power equation, $P = (R_3 + R_4)R_1i^2/(R_1 + R_3)$, and forming the expression for the deflection

FIG. 4.16.—Slide-wire bridge.

$$d = C\sqrt{i_g^2R_g} = \frac{CR_3\sqrt{R_1R_4}\,\alpha\sqrt{P}}{2(R_1 + R_3)(R_3 + R_4)}$$

Thus d is proportional to the fractional unbalance α, which shows that if the resistances cannot be adjusted continuously to achieve an exact balance, it is legitimate to use the galvanometer deflections for the settings on either side of the balance point to interpolate linearly for the next significant figure. Also, d is proportional to the square root of the power input to the bridge and the maximum deflection is obtained by making this equal to the power that the bridge can safely dissipate. On setting the partial derivative of d with respect to R_1 equal to zero it is found that d is a maximum when $R_1 = R_3$. By symmetry d is a maximum subject to the variation of R_4 when $R_4 = R_3$. Thus the optimum sensitivity is achieved when all the resistance arms are equal. In practice the equal-arm bridge has only limited applications, but the general discussion illustrates the considerations involved in designing a bridge for maximum sensitivity. A precaution that should be used in precision work is to reverse the sense of application of the emf. By so doing the effects of stray thermal emfs. in the arms can be eliminated.

The continuous variation that can be obtained with a slide-wire bridge makes it particularly suited to precision measurements. A resistance wire of length $l = l_1 + l_2$ extends between heavy copper terminal blocks, as shown in Fig. 4.16. It is traversed by a sliding contact, the position of which can be read with a vernier and scale. The auxiliary resistances are inserted in the circuit by means of mercury cups in copper terminal blocks as indicated in the figure. The contact and end-correction resistances can be determined in terms of the R's if the r's are removed and their position taken by heavy copper connectors. If ρ is the mean resistance per unit length of the wire, at balance

$$\frac{R_1}{\alpha_1 + \rho l_1} = \frac{R_2}{\alpha_2 + \rho l_2}$$

where the α's are the end-resistance corrections. Reversing the R's

and rebalancing

$$\frac{R_2}{\alpha_1 + \rho l_1'} = \frac{R_1}{\alpha_2 + \rho l_2'}$$

From these equations the α's may be determined. The ratio of the R's should be large to include as great a length of wire as possible. For precision work the wire must be calibrated and the effect of its non-uniformity taken into account. This may be done by means of a small known resistance r equal in magnitude to say $\beta\rho l$, where β is of the order of 0.05. The R's need not be accurately known. Balancing the bridge with r in the position r_1

$$\frac{R_1}{R_2} = \frac{\alpha_1 + \rho l_1 + r}{\alpha_2 + \rho l_2}$$

Moving r to the position r_2

$$\frac{R_1}{R_2} = \frac{\alpha_1 + \rho l_1'}{\alpha_2 + \rho l_2' + r}$$

or $r = \rho(l_1' - l_1) = \rho(l_2 - l_2')$. Knowing the change in position and r, the effective value of ρ in the neighborhood can be deduced. Changing the ratio R_1/R_2 permits the measurement of other portions of the wire. In this way a calibration curve can be drawn. The measurement of nearly equal resistances with the bridge is essentially the reverse of the calibration process. The resistances to be compared are placed in the positions of the r's. Balancing the bridge for one position of the resistances to be compared and then reversing their position and rebalancing yields the equation

$$\frac{\alpha_1 + r_1 + \rho l_1}{\alpha_2 + r_2 + \rho l_2} = \frac{\alpha_1 + r_2 + \rho l_1'}{\alpha_2 + r_1 + \rho l_2'}$$

Adding 1 to both sides, the numerators are equal hence the denominators are equal or $(r_1 - r_2) = \rho(l_2 - l_2')$. Knowing ρ from the calibration and measuring the displacement of the slider the difference between the resistances is determined. This is known as the Carey-Foster method of resistance comparison.

Resistance determinations with a bridge circuit are often used to measure indirectly some other quantity that depends on resistance. As the resistance of a wire is a function of temperature, a thermometer can be constructed on this principle. Platinum is the most suitable metal for the thermometric resistance element as it has a high melting point (1770°C.) and does not oxidize or otherwise deteriorate. The resistance of a platinum wire is given by

$$R_T = R_0(1 + 3.97 \times 10^{-3}T - 0.585 \times 10^{-6}T^2)$$

where R_0 is the resistance at 0°C. and T is the temperature in degrees centigrade, to an accuracy of better than 1°C. from 0°C. to 1000°C. The formula will also yield an accuracy of better than 2°C. from 0°C. to −183°C., the boiling point of liquid oxygen. Thus if R_0 is known and R_T is measured, the temperature of the wire can be deduced. A convenient type of bridge is shown in Fig. 4.17. R_1 and R_1' are approximately equal resistances and the arm R, containing a variable parallel element, can be adjusted over a suitable range to balance the resistance of the platinum wire R_T. In series with R are dummy leads designed to compensate the leads to the platinum element. The latter is generally in the form of a fine wire wound on a mica or ceramic form and preferably contained in an hermetically sealed tube. Since R is proportional to R_T, its setting at balance may be used as a measure of T. In a permanent bridge R may be directly calibrated in terms of T if the highest precision is not necessary. The calibration points chosen are generally: the melting point of ice (0°C.), the boiling point of water (100°C.), and the boiling point of sulphur (444.6°C.). For the highest precision the effects of stray thermal emfs. and resistance changes must be eliminated by techniques that are fully described in thermometric treatises.

FIG. 4.17.—Resistance-thermometer bridge.

Another application of the bridge circuit is in the location of the position of faults in lines or cables. Assume that both ends of the cable are available and that one of the conductors is grounded at some point. One pair of terminals is then connected together and the circuit of Fig. 4.18 is set up. This is a bridge circuit in which the ground is in the battery arm. If the cable is of uniform resistivity, the balance condition is

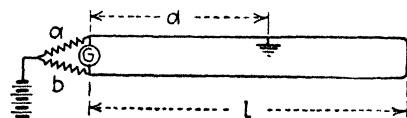



FIG. 4.18.—Fault location in a cable by means of the Murray loop test.

$$\frac{a}{b} = \frac{d}{2l - d} \qquad \text{or} \qquad d = \frac{2}{1 + (b/a)} l$$

Thus a knowledge of the ratio of the arms and the length of the cable determines the distance of the fault from the terminals used. A short circuit between the two conductors can be located by measuring the resistances presented by the two sets of cable terminals if the resistance of the unshorted cable is known. Let R be the resistance of the cable originally, r that associated with the short circuit, and R_1 and R_2 the resistances presented by terminals 1 and 2, respectively, of the faulty cable

with the far terminals open. Then, if l is the length of the cable and d_1 the distance of the short from terminals 1

$$\frac{R_1 - r}{R} = \frac{d_1}{l} \quad \text{and} \quad \frac{R_2 - r}{R} = 1 - \frac{d_1}{l}$$

Eliminating r

$$d_1 = \frac{R_1 - R_2 + R}{2R} l$$

The resistance associated with the short can also be found for

$$r = \frac{R_1 + R_2 - R}{2}$$

The application of the bridge circuit to shorted cables is limited, for frequently the shorting resistance r is not constant and if it is much greater than the R's, the measurements cannot be made accurately.

FIG. 4.19.—Measurement of a high resistance by the rate of discharge of a condenser.

The simple bridge circuit is not well adapted to the measurement of resistances greater than about 10^7 or less than 10^{-1} ohm. One method of measuring a high resistance is to place it in series with a known emf. and a calibrated galvanometer. The galvanometer deflection is a measure of the current and the ratio of the emf. to this current is the resistance of the circuit. A method that can be used for still higher resistances is to measure the rate at which a condenser discharges through the resistance. The circuit is illustrated in Fig. 4.19. The potential difference across the condenser is measured with an electrometer and by the fundamental circuit theorem

$$V_C + V_R = 0$$

since no emf. is present. V_C is equal to q/C and $V_R = iR(i = dq/dt)$, where q is the charge on the condenser at a time t. Thus

$$\frac{dq}{dt} + \frac{q}{RC} = 0$$

Integrating this equation:

$$\log_e q = -\frac{t}{RC} + \text{const.}$$

or, if $q = q_0$ at, say, $t = t_0$, the constant is determined as $\log_e q_0$ and

$$\log_e \frac{q}{q_0} = -\frac{t}{RC} \quad \text{or} \quad q = q_0 e^{-\frac{t}{RC}} \tag{4.25}$$

If the electrometer deflection is proportional to V_C, it is also proportional to q and the ratio of any two deflections separated by a known time interval will determine the quantity RC. If C (which includes the electrometer capacity) is known, R can be found. A greater accuracy can generally be achieved by plotting the log of the ratio of the deflection to the initial deflection against the time. The negative slope of the straight line obtained is the reciprocal of the product RC.

The simple bridge circuit is not suited to the accurate measurement of very low resistances because of the errors introduced by the network junctions which are effectively included in the bridge arms. These are generally of the order of 0.0001 ohm and become important when the resistance to be measured is less than 0.1 ohm. One circuit for the elimination of terminal and contact resistances is the Kelvin double bridge which is shown in Fig. 4.20. X is the unknown resistance and N is the standard with which it is to be compared. The resistance of the junction between these two is r. A current of several amperes is sent through these resistances in series. Its value can be varied by the external resistance R' and measured by the ammeter A. The reading of A need be noted only if the variation of the resistance X as a function of the current carried by it is to be studied. The resistances R_a, R_b, R_c, and R_d constitute the bridge proper and the galvanometer is used as a null instrument to determine the balance condition. If the ratio of the sum of the potential drops in X and R_a to that in R_c is equal to the ratio of the sum of the potential drops in R_b and N to that in R_d, the galvanometer terminals will be at the same potential and the bridge will be in balance. If i is the current through X and N, i_1 that through R_c and R_d, and i_2 that through R_a and R_b this condition may be written

FIG. 4.20.—Kelvin double bridge.

$$\frac{Xi + R_a i_2}{R_c i_1} = \frac{Ni + R_b i_2}{R_d i_1}$$

This condition is seen to be fulfilled for any values of the currents if

$$\frac{X}{N} = \frac{R_a}{R_b} = \frac{R_c}{R_d} \tag{4.26}$$

Equation (4.26) is the balance condition for the bridge and it is seen to be independent of the junction resistance r which is the desired result. The R's composing the bridge are sufficiently large that the junction resistances in series with them can be neglected.

There are two general methods of using this bridge. In one the R's are fixed in the proper ratio and the standard resistance N is a heavy rod of known resistance per unit length. The junction of say R_b and N is made by means of a clamp that can be moved along the rod. The value of N is obtained from the length of the rod included in the bridge arm and X is determined from Eq. (4.26). In the other method N is a fixed standard, R_b and R_d are equal and constant, and R_a and R_c are variable but always maintained equal to one another. This may be accomplished by double switches that change these resistances by equal amounts. X is then given by Eq. (4.26) as the product of N and the ratio of the variable to the fixed resistances. The second method of use is better adapted for the comparison of standard resistances and is more precise than the first. Thermal emfs. may be eliminated by reversal of the battery. The sensitivity increases with the square root of the power supplied. This bridge is generally used for measuring resistances in the range from 10^{-1} to 10^{-3} ohm. For lower resistances a potentiometer method is generally employed.

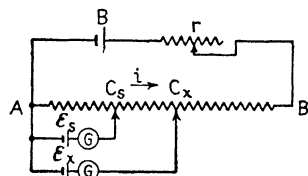



FIG. 4.21.—The potentiometer principle.

4.7. The Potentiometer.—The term potentiometer is used to refer in general to any resistance element with one or more taps or sliding contacts if both terminals of the resistance are available. It is also used more specifically to designate a precision resistance of this type which together with pertinent accessories is used for the comparison of potential differences. The general potentiometer principle is widely used in all types of electrical circuits, and the precision potentiometer is probably the most important instrument in the field of electrical measurements. The principle of operation of the potentiometer, which is extremely simple, is illustrated in Fig. 4.21. A battery is connected in series with a variable resistance r and the potentiometer resistance R which extends between the points A and B. The potential difference between A and B is iR, where i is the adjustable but in general unknown current flowing in this circuit. If R_s is the resistance of the portion of the circuit between A and C_s and R_x that of the circuit from A to C_x and if V_s and V_x represent the potential differences between A and C_s and A and C_x, respectively,

$$\frac{V_x}{V_s} = \frac{iR_x}{iR_s} = \frac{R_x}{R_s} \quad \text{or} \quad V_x = \frac{R_x}{R_s}V_s \tag{4.27}$$

Thus, if V_s and the resistance ratio are known, V_x is determined. Now, if one of the lower circuits, say the one containing $\mathcal{E}_s$, is established and the galvanometer shows no deflection, there is no potential difference

across the galvanometer terminals and V_s is equal to the potential difference across the terminals of the cell. Likewise, if the galvanometer shows no deflection, there is no current through the cell and the potential difference between its terminals is the open-circuit potential difference which is the emf. of the cell. If the second of the lower circuits of the figure is established as well and that galvanometer shows no deflection $\mathcal{E}_x$ must equal V_x. Hence, if the resistance from A to B is so calibrated that the ratio of the resistances corresponding to the settings C_s and C_x is known, the emfs. $\mathcal{E}_s$ and $\mathcal{E}_x$ can be directly compared by means of Eq. (4.27) without a knowledge of the absolute magnitudes of these resistances or of the current flowing through them, provided the latter is the same when the two measurements are made. The fact that no current is drawn from the emf. under measurement is of great importance, for the effect of any resistance in this circuit is thus eliminated.

The standard $\mathcal{E}_s$ in terms of which ordinary laboratory measurements are made with this instrument is generally the Weston standard cell (Sec. 6.2). This is a specially prepared type of voltaic cell which is characterized by an extreme constancy of emf. under various conditions of operation and over long periods of time. The emf. developed by the cell is approximately 1.0187 volts at 20°C. with a negative temperature coefficient of about 4×10^{-5} per degree centigrade. The accurate emf. of a particular cell is given by the calibration accompanying it. The constancy of the emf. of a properly aged cell can be relied upon to 1 part in 10^5 over a period of years after calibration if carefully used. *The current drawn from the cell should never be greater than 1 μa. for if larger currents are drawn, the calibration of the cell is apt to be changed.* In a potentiometer circuit a protective resistance is included in series with the cell and galvanometer in order that this current will not be exceeded in obtaining a preliminary balance. The standard cell is, of course, merely a secondary standard of potential which is calibrated in a national standardizing laboratory in terms of the absolute standards of current and resistance. Its permanence and the simplicity of the comparison technique makes it one of the two important electrical standards in ordinary laboratories. The other important standard is, of course, that of resistance, and in conjunction the two may be used for the calibration of current- and power-measuring instruments.

Many different forms of precision potentiometer have been designed, but of these only one type of general utility will be described. This is known as the Leeds and Northrup Type K. The circuit is shown schematically in Fig. 4.22. The principal resistance train, corresponding to AB of Fig. 4.21, is composed of the resistances associated with the three dials of the figure. The taps to the central dial are taken off at equal intervals of 5 ohms, and the dial at the right which represents a

slide wire on a drum also has a resistance of 5 ohms. Thus the resistance range of the drum is equal to that between taps of the central dial. The train is designed to carry 0.02 amp. so that when this standard current is established, the potential difference corresponding to one step of the central dial is 0.1 volt. The slide-wire scale has 1,000 main divisions and may be read to 0.1 division so that the instrument is capable of measuring an emf. of 1 volt with a precision of 1 part in 10^5. No variable contact resistances occur in the potentiometer chain. The switch

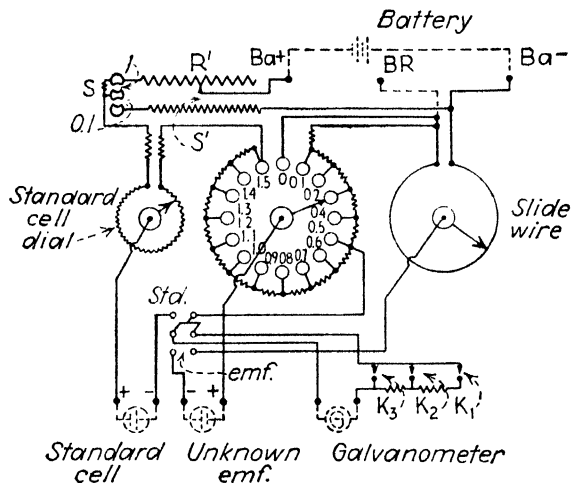



FIG. 4.22.—Leeds and Northrup Type K potentiometer.

contacts are in the galvanometer circuit and as these carry no current at balance, no error is introduced. The terminal blocks associated with the resistances s and s' in the upper left corner are designed for the accommodation of a plug which is normally in the upper position. When it is moved to the lower position, s is in series with the train and s' in shunt with it. These resistances are in such a proportion to that of the train that this change reduces the current through the train to exactly $\frac{1}{10}$ of its original value without changing the resistance presented to the battery circuit. This reduces the potential difference appearing between the dial arms by a factor of 10 and permits the measurement of a potential in the neighborhood of 0.1 volt with a precision of about 1 part in 10^5, *i.e.*, to about 1 μv. The left-hand dial is for use only in conjunction with the standard cell. It is variable over a very small interval to allow for the variation in emf. between individual standard cells. The potential difference against which the standard cell is compared is derived from this switch arm and one of the taps of the central dial.

The procedure of use is as follows: A battery composed of two dry or storage cells is connected in the appropriate sense to the terminals Ba_+ and Ba_-. The standard cell and galvanometer are connected to the terminals provided for them, the proper sense of connection being of course observed in the case of the cell. The scale-changing plug is placed in position 1 and the standard-cell dial is set at the proper position for the emf. of the standard cell being used, with any correction necessary for the ambient temperature. The double-pole double-throw switch is set in the position marked "standard" and the resistance R' is varied until on tapping key 1 no galvanometer deflection is observed. The final adjustment of R' is reached by using key 2 and then key 3, which successively eliminate the protective resistances in series with the cell and galvanometer. This procedure adjusts the current through the potentiometer train to exactly its proper value so that the dials of the instrument read directly in volts. The potentiometer is then said to be calibrated. The double-pole double-throw switch is then set in the position marked "emf." and an unknown potential difference may be applied to the terminals marked "emf." and compared with that between the contacts traversing the two principal dials. In this procedure the settings of these dials are varied and the keys 1, 2, and 3 tapped successively as the null deflection of the galvanometer is approached. When the adjustment is complete, the value of the unknown emf. is read from the dials. An emf. greater than 1.6 volts cannot be measured with this instrument directly. If the emf. used is less than this but a balance cannot be achieved, the emf. terminals have been applied to the instrument in the wrong sense and must be reversed. In making a series of measurements the potentiometer must be recalibrated at frequent intervals as the current through the resistance train is apt to change slowly with the time.

The potentiometer finds application in almost every field of electrical measurement. It is indispensable in electrochemistry for measuring the emf. developed by various voltaic cells, as important chemical quantities can be deduced from these measurements. Most of these emfs. lie within the ordinary range of the instrument. Photovoltaic and thermoelectric emfs. are generally considerably smaller than those developed by chemical cells and for these the potentiometer is used in its lower range. In making these measurements it must be remembered that the instrument must be returned to its normal range for calibration. These applications are discussed in more detail in Secs. 6.2 and 6.3. In order to use the potentiometer for measuring emfs. above 1.6 volts a greater potential drop must occur in the resistance train. It is not advisable to apply 10 times the battery potential to the Leeds and Northrup type of instrument, performing the calibration on the 0.1 scale

and returning then to the normal scale for measuring emfs. up to 16 volts, because of the greatly increased power that must be dissipated by the potentiometer. In normal operation the instrument dissipates only a few hundredths of a watt, while in such an arrangement it would have to dissipate several watts, which would endanger the resistances. In a higher resistance potentiometer such as the Queen or Wolff type the dissipation of power is less for the same applied potential ($P = V^2/R$) and this method of increasing the range is feasible.

A voltmeter can be calibrated directly with a potentiometer if its range lies below 1.6 volts. An auxiliary battery and variable resistance arranged as a simple potentiometer is used to apply an adjustable

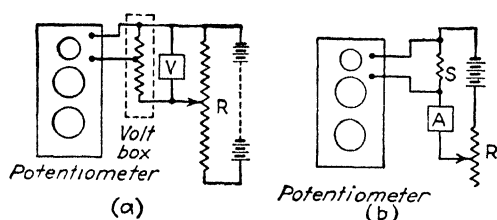



FIG. 4.23.—Instrument calibration with a potentiometer. (a) Voltmeter calibration. (b) Ammeter calibration.

potential difference to the meter terminals. To these terminals are also connected the emf. terminals of a precision potentiometer. The scale reading of the meter is noted and also the potentiometer reading at balance. By varying the potential difference applied to the meter terminals and noting the scale and potentiometer readings the scale of the meter can be calibrated. If the scale of the meter lies beyond the ordinary range of the potentiometer, an auxiliary precision resistance known as a "volt box" must be used. This is a resistance which is tapped at accurately known fractions of its total value. These fractions are generally 0.1 and 0.01. The circuit is illustrated at the left in Fig. 4.23. The potential difference across the meter terminals is also applied to the volt-box terminals and, say, the fraction 0.01 of it is applied to the potentiometer terminals. Thus at balance the potentiometer indicates 0.01 of the potential difference across the meter terminals. Therefore a meter with a 150-volt scale can be calibrated in this way. The volt box, of course, draws a current from the external circuit and hence is not adapted to the measurement of emfs.

An ammeter can be calibrated with a potentiometer and a standard resistance. The circuit is shown at the right in Fig. 4.23. The standard resistance S should be of the order of V'/i', where V' is the maximum potential that can be measured by the potentiometer on one range or the other and i' is the current for full-scale deflection of the meter.

For example, if the meter shows a full-scale deflection for 15 amp., S should be a 0.1-ohm standard. The current through A can be adjusted by varying R, and its value is given by the potential difference appearing across S, as measured by the potentiometer, divided by the value of S. Thus to calibrate the meter scale the scale and potentiometer readings are noted for a series of settings of R.

An unknown resistance can be compared with a standard by connecting them in series with an emf. The potential drop across each resistance is measured with a potentiometer, and since the same current flows through both the resistances these are in the ratio of the potentials measured across their terminals. The power delivered to a load can also be measured in an obvious way by making two potentiometer measurements; one across a standard resistance in the line and the other using a volt box across the line.

Problems

1. How many electrons pass per second through the filament of a 100-watt lamp if a potential difference of 110 volts exists between its terminals?

2. Fifty 100-watt lamps are operated in parallel on a 110-volt line. If the line has a resistance of $\frac{1}{20}$ ohm, what fraction of the power supply by the generator is lost?

3. An electric immersion heater, which draws 6 amp. from the 110-volt line, will bring 1 liter of water to a boil in 10 min., starting from a temperature of 20°C. What is the efficiency of the heater (per cent of the energy supplied that goes to heating the water)?

4. The anticathode of an X-ray tube consists of a hollow water-filled cylinder. If 1 liter of water is evaporated per hour when the tube is operating at a current of 10 ma., what is the potential difference across the tube? (Latent heat of vaporization of water is 539 cal. per gram.)

5. The copper wire of a circuit is 2 mm. in diameter and it is protected by a fuse wire 1 mm. in diameter. Taking the constants of the copper and fuse wire as

	Copper	Fuse wire
Resistivity	1.7×10^{-6}	3.5×10^{-5} ohm cm.
Density	8.9	12 gm./cm.^{3}
Specific heat	0.091	0.051 cal./gm. °C.
Melting point	.	185°C.
Heat of fusion	42	6 cal./gm.

and assuming room temperature to be 20°C., how long will it take for the fuse wire to melt for a short-circuit current of 25 amp. and what will then be the temperature of the copper wire? (Neglect the loss of heat by conduction, etc.)

6. Three resistances of 1, 3, and 5 ohms, respectively, are connected in parallel and the group placed in series with a resistance of 7 ohms, a battery of 3 ohms internal resistance, and an emf. of 15 volts. Calculate the current flowing and the rate of generation of heat in the separate elements.

7. A wire of uniform resistance r per unit length is bent into the form of an equilateral triangle. The wire is soldered together at the apex and wire of the same material used to join the mid-points of the sides. If the sides of the triangle are of length $2l$, calculate the resistance offered by the network to a current that enters at one apex and leaves at another. Calculate the currents in the separate branches if the applied emf. is V.

8. Eight equal resistances are arranged in the form of a square with diagonals connected at the mid-point. If the resistance of each wire is r, calculate the resistance between (*a*) opposite corners, (*b*) adjacent corners, (*c*) a corner and the center of the square.

9. Ten unit elements of resistance r are arranged in the form of a ladder with four rungs. Calculate (*a*) the resistance between terminals at one end, (*b*) the resistance between diagonal terminals, (*c*) the resistance between neighboring central terminals on the same side of the ladder.

10. Six wires each of resistance r are connected together, forming the edges of a regular tetrahedron. Find the resistance presented by the network (*a*) at any two corners, (*b*) at the mid-points of opposite sides. Find the power dissipated in the separate arms in both cases.

11. The exponential function is defined by the series:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

and the error made in stopping at the nth term is of the order of $e\ /n!$. Given 160 6-ohm resistances, how would they be combined in a network to which a current of e amp. would flow when a potential difference of 1 volt is applied to it? By how much may the value of the current be in error?

12. A wheel of radius l has a wire of resistance R' per unit length wrapped around its rim and a piece of the same wire runs along one spoke connecting the rim and axle. Find the resistance between the axle and a sliding contact on the rim as a function of the angular rotation of the wheel. Assuming that a potential difference V is applied between the contact and axle, plot the current that will flow as a function of the angle for two complete rotations.

13. Assuming the wheel of the previous problem has wires connecting the rim and axle at three equally spaced spokes, calculate the resistance between the contact and axle as a function of the angular rotation and find its maximum and minimum values.

14. A battery composed of cells of emf. $\mathcal{E}$ and internal resistance R_e is connected in series with a relay of resistance R_r and the wheel of Prob. 12. Assuming that the relay operates at a current i, how many cells must be placed in series for the relay to be closed during half a revolution?

15. How great must be the insulation resistance between the terminals of a megohm resistance box in order that it shall not be in error by more than 1 per cent?

16. How low must the junction resistances of a 0.1-ohm shunt be in order that the total resistance the shunt introduces shall be its nominal value to within 1 per cent?

17. A dry cell shows an emf. of 1.579 volts as measured by a potentiometer. When a 1000-ohm voltmeter is connected to its terminals, the meter reads 1.52 volts. What is the internal resistance of the cell?

18. The emf. of a cell is measured by a potentiometer to be 1.42378 volts. It is then shorted with a resistance of 5×10^5 ohms and the potential difference across its terminals is found to be 1.42319 volts. What is the internal resistance of the cell and with what accuracy has it been measured if the last figure of the potential may be in error by half a unit?

19. A resistance R is placed across the terminals of a cell having an emf. of 1.5 volts and an internal resistance of 1.7 ohms. Calculate and plot the power delivered to R as a function of its value from 1 to 10 ohms. Plot also the current, the potential difference across R, the total power, and the fraction dissipated in R.

20. A set of n cells, each of emf. e and internal resistance r, are connected together in a circuit in such a way as to deliver the maximum power to a load of resistance R. If there are p parallel groups of s cells in series, show that s and p are the integers nearest $\sqrt{nR/r}$ and $\sqrt{nr/R}$, respectively. Show that the power delivered to R under these circumstances is half the total power expended or $e^2n/2r$.

21. An electrometer of capacity 5×10^{-11} farad is connected across the terminals of a condenser with a capacity of 10^{-10} farad. The combination is charged to such a potential that the deflection of the electrometer is d. After 100 sec. the electrometer deflection is found to be half its original value. What is the leakage resistance of the insulators in the circuit? The condenser is then recharged to the original value and a resistance is placed across its terminals. After 15 sec. the electrometer deflection is half its initial value. What is the value of the resistance?

22. A galvanometer is placed in the position of R_2 of the bridge circuit of Fig. 4.15 and the galvanometer there shown is replaced by a tapping key. Show that if the galvanometer deflection is not affected by tapping the key, the resistance of the galvanometer is R_1R_4/R_3.

23. Show that Kirchhoff's laws imply that the currents so distribute themselves in a passive network that the rate of generation of heat is a minimum.

24. A pair of No. 10 copper-clad steel telegraph wires, having each a resistance of 13.2 ohms per mile, form a line 10 miles long. A ground occurs on one of the wires and on setting up the circuit of Fig. 4.18 balance is obtained when $b = 75$ ohms if $a = 100$ ohms. How far is the ground from the terminals used?

25. The two wires of the previous problem are short-circuited at some point. The resistance presented by terminals 1 when the others are open is 120 ohms and that presented by terminals 2 with terminals 1 open is 158 ohms. Where is the short and what is its effective junction resistance?

26. A potentiometer train has a resistance R and the current is sent through it by a battery of emf. $\mathcal{E}'$ and negligible internal resistance. It is used with a galvanometer of resistance R_g to measure an emf. $\mathcal{E}$ generated in an element of internal resistance R_e. Show that the current through the galvanometer is given by

$$i_g = (\mathcal{E} - x\mathcal{E}')[R_g + R_e + x(1 - x)R]^{-1}$$

where x is the friction of R included in the galvanometer circuit. Assuming $\mathcal{E}$ and R_e to be fixed, find the optimum values of the circuit parameters if the galvanometer deflection is proportional to i_g and $\sqrt{R_g}$.

27. Design an attenuator of the type of Fig. 4.9 to attenuate in steps of 2 db. from 0 to 50 db and work between resistances of 600 ohms.

28. Show that if the product of the load resistance and the value of g of a cable is greater than 1, the hyperbolic cotangent would replace the hyperbolic tangent in Eq. (4.21), where then $\coth \phi = R_lg$, and that sinh would appear for cosh and cosh for sinh in Eq. (4.22).

29. Show from Eq. (4.22) that if R', G', and l are small enough so that the second power of αl can be neglected in comparison with unity,

$$V_l = \frac{V_0}{1 + \dfrac{R'l}{R_l}}$$

30. A cable 100 km. long is composed of two No. 14 copper conductors each of which has a resistance of 8.285 ohms per kilometer. If the insulation resistance is 10^{12} ohms per meter, show that when a potential difference of 100 volts is established across the terminals at one end, a current of 38.5 ma. will flow through a 1,000-ohm resistance across the far terminals.

31. The total resistance of a concentric conductor cable is 10^{-4} ohm per meter and the leakage resistance of the insulation is 10^{10} ohms per meter. Calculate the current that will be drawn from a 100-volt battery placed across the two terminals at one end of a 10,000-km. cable (*a*) if the far end is open, (*b*) if the far end is short-circuited. Find the current that would flow through a 1,000-ohm resistance across the far end.

32. A line supported by m poles has a total resistance of R ohms per section between poles and a leakage resistance of r ohms at a pole. Show that if V_0 is the potential difference applied to one end of the line, the potential difference between the lines at the nth section is

$$V_0 \frac{\sinh [(m - n)\sqrt{R/r}]}{\sinh [m\sqrt{R/r}]}$$

and the current flowing in the line there is

$$\frac{V_0}{\sqrt{Rr}} \frac{\cosh [(m - n)\sqrt{R/r}]}{\sinh [m\sqrt{R/r}]}$$

if m and n are large and the far end is short-circuited. Find the values of these quantities if the far end is open.

33. Assume that the shunt equations for a network have been solved and the quantities (B_{il}/D') determined. Show that if additional currents from some external source equal to i_l at junction l are injected into the network, the potentials of each of the $(J - 1)$ junctions will be increased by the amount

$$\delta V_i = \sum_{l=1}^{l=(J-1)} \frac{B_{il}}{D'} i_l$$

34. An accelerometer consists of a metal cube of mass M that is supported between two thin sheets of spring steel affixed to opposite faces of the cube and in turn supported at opposite pairs of edges by a rigid framework. The block can thus move elastically only in the direction normal to the spring supports. The restoring force per unit displacement of the spring is k, and the displacement in either direction is limited by stops on the supporting framework to prevent undue amplitude of motion of the block. Two pairs of wires, each of the same resistance, are connected between the block and framework across the gaps between the block and the stops. These are each of length l in the direction of permitted motion and are connected to form the arms of a Wheatstone bridge, the wires crossing the same gap forming opposite bridge arms. Assuming that the input and output circuits are matched and that the ratio of fractional change in wire resistance to its elongation per unit length is C, show that the galvanometer current measures the component of acceleration of the framework a normal to the springs through the relation

$$i_g = -\frac{CMi}{2lk} a$$

where i is the current supplied by the bridge battery.

CHAPTER V

NONOHMIC CIRCUIT ELEMENTS AND ALTERNATING CURRENTS

5.1. Introduction.—In the preceding discussion of the conduction of electricity it has been assumed that Ohm's law has applied. This is equivalent to assuming that the current density in the conducting medium is proportional to the electric-field strength

$$\mathbf{i}_v = \sigma \mathbf{E} \qquad \text{or} \qquad V = Ri \tag{5.1}$$

where σ and R are constants. Metallic conduction follows this law very accurately and metals as a class are the typical linear or ohmic conductors. The great majority of other solid materials are ohmic to a very good approximation, and the conduction of electricity in liquids, with certain restrictions, also follows this law. Gaseous conduction, however, cannot be represented even approximately in this way. There are also many interesting and important instances of solid conduction for which Ohm's law does not hold. These are known as *nonlinear conductors* and circuit elements composed of such conductors are called *nonlinear circuit elements*. In terms of the simple conduction concepts of Sec. 3.3 nonlinearity implies that either the number of electrons available for conduction or their mobility is influenced directly or indirectly by the applied electric field.

In general, the current that flows through an element forming part of an electric circuit depends not only on the potential difference across its terminals but on other physical parameters as well, such as the temperature, state of strain, etc. In effect, R of Eq. (5.1) is a function of the physical condition of the conductor even for metals that are generally considered as ohmic conductors, and a unique relation will exist between i and V only if these parameters are unaltered. As an example consider that a copper wire is stretched tightly between two terminals and that a potential difference V is maintained between these ends. Electrical energy is transformed into heat at the rate iV and this increases the temperature of the wire. As the temperature rises the apparent resistance (V/i) increases and as a secondary effect the thermal expansion of the wire decreases the tension which will in general also affect the apparent resistance. Thus the measured ratio of V to i will not be constant and the wire under these conditions is essentially a nonlinear conductor.

It is convenient to divide nonlinear elements into two types: ***intrinsic***

and *contingent.* By an intrinsically nonlinear element is meant one in which a unique single-valued nonlinear functional relation exists between i and V. By a contingently nonlinear element is meant one in which the relation between i and V is not unique but depends upon other parameters, which may include the previous history of the element, time, temperature, state of strain, or any other physical factors that affect the relation between i and V. Examples of the intrinsic types are certain ceramic compounds, boundary layers between metals and semiconductors, and also vacuum tubes under ordinary conditions of operation. The contingent type is represented by any element for which i is not a uniquely determined function of V. Under specific conditions of operation the values of i and V may themselves determine the values of all other parameters affecting the relation between them. As an instance of such a case consider an otherwise ohmic element for which R is a function of the temperature. In a static or dynamic steady state the temperature of the element is determined by the power generated in it and the rate of loss of heat to its surroundings. The former is determined by i and V; and hence if the latter is constant, the element resembles an intrinsically nonlinear element, although if i and V are functions of the time, the effective relation between them may not be single valued.

For the intrinsically nonlinear element it is clearly possible to represent the relation between i and V graphically. In certain cases approximately analytical relations can also be found. This analytical or graphical functional relation between i and V is known as the *characteristic* of the element or, more strictly, as the *static characteristic*, as it is assumed that points on the curve represent steady equilibrium values of i and V. In the case of a contingent element these values can be attained only after a considerable time. If V or i are functions of the time, the resulting current or potential difference is uniquely determined by the static characteristic in the case of an intrinsic element. This is not, in general, true for the contingent element. If the fluctuation is very slow, the static characteristic may be traversed, but in general an entirely different curve will be traced on the i-V diagram. If the alternating component is periodic, a steady dynamic state will be achieved, and the curve representing this relation between i and V is known as the *dynamic characteristic.* This curve is not in general single valued, and its shape will depend markedly upon the amplitude of the alternating components of i and V and upon the frequency of alternation. For a nonlinear element the concept of resistance loses much of its significance. However, the *static* or *apparent resistance* is defined as the ratio V/i for points on the static characteristic by analogy with Eq. (5.1). In general, of course, it has a different value for every point on the characteristic. If the alternating components of V and i are small, the value of dV/di, which

is the slope of the characteristic, assumes particular significance. It is given the designation *dynamic resistance,* and it will be widely used in the discussion of vacuum-tube circuits. In the sections immediately following, the static characteristics of a few representative nonlinear elements will be discussed. After the fundamental concepts of alternating potentials and currents have been introduced, the discussion will be extended to include dynamic characteristics.

5.2. Intrinsically Nonlinear Elements.—It is convenient to consider intrinsically nonlinear elements in two groups: those for which $i(V) = -i(-V)$, which are known as *symmetrical elements,* and others for which this relation does not hold, which are *asymmetrical elements.* In the case of a passive element that contains no source of energy the static characteristic lies in the quadrants for which i and V are of like sign, and it passes through the origin. If a source of emf. or current is part of the element itself, the curve may obviously be displaced by a corresponding amount along one axis or the other. Although the most familiar instances of intrinsically nonlinear circuit elements will later be met in the discussion of vacuum tubes and gas discharges, there are important types of solid conductors which exhibit pronouncedly nonohmic characteristics.

An example of a symmetrical intrinsically nonlinear element is a black ceramic material formed by heat treatment of a mixture of clay and carbon known as *thyrite.*[1] Its characteristic is represented to a very good approximation by an equation of the form

$$i = AV^a \tag{5.2}$$

Here the constant a has a value very close to 3.5 for most samples, and A, of course, depends on the length and cross section of the specimen. The most convenient way to plot such a function is in terms of the logarithms of i and V as shown in Fig. 5.1. On such a plot the curves of constant resistance and constant power are straight lines representing constant differences and sums, respectively, of the coordinates.

It is seen that an increase in voltage by a factor of 10 produces more than a 3,000-fold increase in current. This property of thyrite makes it particularly useful in protecting electrical equipment from overvoltages. For instance, a transmission line may be grounded through a piece of thyrite which passes a current of the order of a milliampere under normal conditions. If the line is struck by lightning and the potential rises from 100,000 to 1,000,000 volts, the excess charge which produces this dangerous condition is drained off at an initial rate of about 3 amp. dropping off again to the value of 1 ma. as the normal condition of the line is restored. Similarly, it will be seen later that if a circuit containing an iron-cored

[1] BROWNLEE, *Gen. Elec. Rev.* **37**, 175, 218 (1934).

coil of wire, which is carrying a current, is suddenly opened, a large potential difference will instantaneously occur across the terminals of the coil. This voltage surge which endangers the insulation may be greatly reduced by connecting a piece of thyrite permanently across these terminals, in parallel with the coil. It draws a negligible current under normal circumstances but supplies a low-resistance path between the coil terminals in

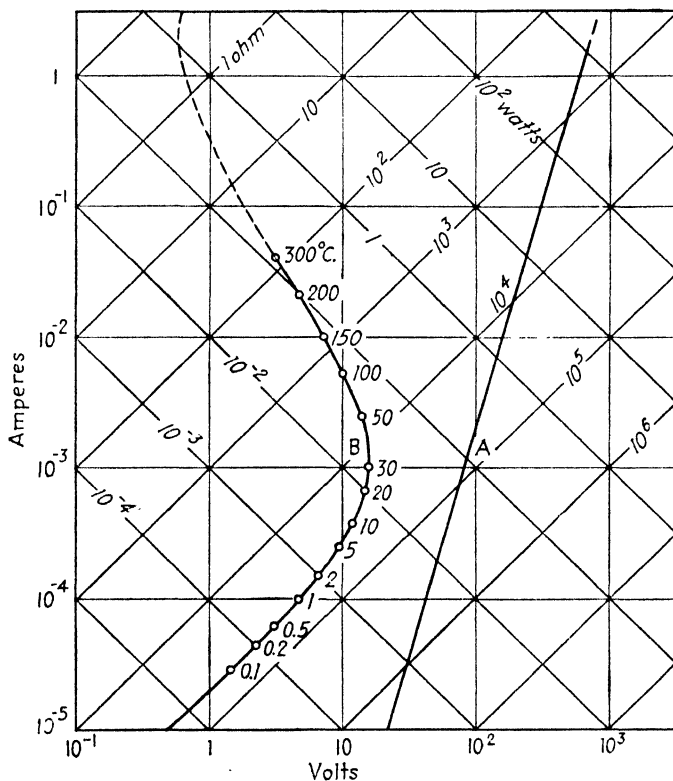



FIG. 5.1.—A, thyrite disk, 6 in. diameter, $\frac{3}{4}$ in. thick. $i = 2.13 \times 10^{-10}V^{3.5}$. B, thermistor, representative static characteristic. Numbers designate ambient temperature in degrees centigrade at the points. Eqs. 5.5 and 5.6 parameters: $R_0 = 50{,}000$ ohms, $\beta = 5 \times 10^{-4}$ watt/degree, $B = 3{,}900$ per degree. The dashed portion of the curve exceeds reversible power levels.

the presence of the voltage surge. Many uses for these and other nonlinear conductors will develop in connection with alternating currents.

Instances of asymmetrical intrinsically nonlinear elements are contacts between metals and *semiconductors.* A semiconductor is a material such as copper oxide, zinc oxide, galena, silicon, germanium, or selenium. These will be discussed in more detail in Secs. 5.4 and 6.3. At this point the interest centers in the contacts between such materials and metals. These junctions have the property of exhibiting different resistances for the two possible senses of current flow. If the current is kept small, the

resistance is in general much greater for one sense of flow than for the other, and these elements form the basis for one class of alternating-current rectifiers. The characteristic may be represented to a fair degree of approximation for small currents by the expression

$$i = A(e^{bV} - 1) \tag{5.3}$$

where A and b are constants. A representative characteristic is shown in Fig. 5.13. The constant A depends on the size, shape, and nature of the material as well as the temperature through a factor of the form $e^{-\frac{\varphi}{kT}}$ where k is Boltzmann's constant (1.37×10^{-23} joule/°C.) and φ is the difference between the work functions of the metal and semiconductor (Sec. 6.3). Thus these junctions are intrinsically nonlinear elements only if the temperature remains sensibly constant. Many of these metal-semiconductor junctions exhibit photo effects. That is, changes in the current for a constant applied emf. or the appearance of a potential difference at the boundary layer when there is no applied emf. are observed when the element is illuminated.

Asymmetrical junctions of this type are known as *rectifying junctions*. They can be divided into two classes: those used for handling relatively large amounts of power and those used as rectifiers at low power levels. When copper oxide or selenium units are prepared with relatively large areas of contact, currents of many amperes can be safely handled. In such applications, units are generally arranged in series-parallel arrangements such as those shown in Fig. 5.18. As these rectifying junctions have appreciable conductance in the inverse direction, especially for large inverse voltages, it is often advantageous to place several units in series in order that no one unit be required to withstand too large an inverse voltage. Large-area units can be used only for audio and power frequencies because of the capacity associated with the large area of contact. Rectifiers that are used as communication circuit elements usually have small areas of contact and often consist of a tungsten cat's whisker in contact with a piece of silicon or germanium. Such units have very small areas of contact and can be used at frequencies as high as 10^{10} cycles per second.

5.3. Circuits Containing Both Ohmic and Nonohmic Elements.—Circuits containing nonlinear elements are in general much more difficult to analyze than are ordinary ohmic circuits. Of course, Kirchhoff's laws are applicable, but the equations resulting are often insoluble by algebraic methods. The general circuit laws of Sec. 4.3 no longer apply to the complete network. They do, of course, continue to apply to the ohmic portions; and if only one nonlinear element is contained in the circuit, the equations can always be put in the form of an emf. in series with the

nonlinear element and an ohmic resistance. In consequence it is of interest to consider methods of obtaining the relation between currents and potential differences for simple combinations of ohmic and nonohmic elements.

Although certain cases can be handled analytically, it is generally necessary to resort to graphical methods. Consider a nonlinear element having the characteristic $V = V(i)$ represented by S in Fig. 5.2 in series with a resistance R. The current flowing through both of these elements is necessarily the same, say i_1. Hence if a straight line making an angle $\theta = \tan^{-1} R$ with the current axis is drawn through the point on S having the ordinate i_1, the difference in potential between its intersection with the V axis and the abscissa of the point having the ordinate i_1 is the potential difference appearing across R. Thus the point of intersection gives the current that flows through the series combination when a potential difference V_2 is applied to it. This type of plot is widely used in the analysis of vacuum-tube circuits, and the composite characteristic $V_2 = V_2(i_1)$ can be obtained by a series of choices of either i_1 or V_2. A somewhat more convenient method of deriving the composite characteristic graphically is shown in Fig. 5.3. The curve S represents the characteristic of the nonlinear element, and the straight line R is drawn through the origin at an angle $\theta = \tan^{-1} R$ with the current axis. If R and S are in series, the currents through them are the same, and hence adding algebraically the abscissas of the curves for each value of the ordinate yields the series composite characteristic as given by C. If the elements were in parallel, a similar argument shows that the composite parallel characteristic is given by adding together the ordinates of the curves for each abscissa. This curve, of course, lies on the opposite side of R and S from C. By iteration of this method the composite characteristics of any combination of ohmic and nonohmic elements can be obtained.

FIG. 5.2.—Graphical treatment of a simple series circuit containing a nonlinear resistance.

There is one interesting and useful circuit containing two nonlinear elements for which the analysis is fairly simple. This is the symmetrical bridge circuit of Fig. 5.4. Let the areas marked T be similar thyrite samples, though with the appropriate modifications in the analysis they could be lamp filaments or other nonlinear elements. The resistances R comprise the other two arms of the bridge, R_L is a load resistance,

and V is the potential applied to the bridge. Kirchhoff's laws yield the following equations:

$$i_T = i_L + i_R$$
$$V = Bi_T^b + Ri_R$$
$$Ri_R = Bi_T^b + R_L i_L$$

Eliminating i_R, the following parametric equations are obtained relating i_L and V:

$$i_L = \frac{Ri_T - Bi_T^b}{R + R_L}$$
$$V = Bi_T^b + Ri_T - Ri_L$$

On assigning a series of values to i_T the curve of Fig. 5.5, giving the relation between i_L and V, is obtained. The most interesting point

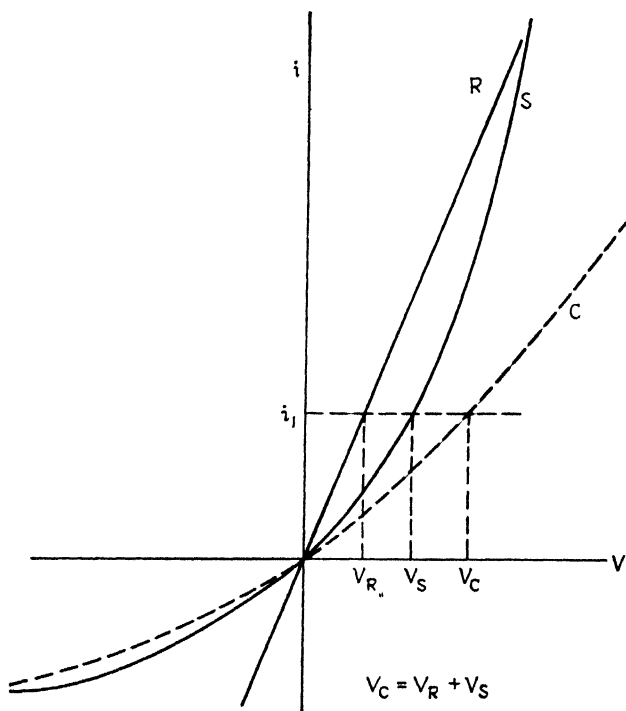



FIG. 5.3.—Graphical determination of a composite series characteristic.

about this curve is the maximum which occurs at $V = V'$. This maximum is determined by the condition $di_L/dV = 0$. The same condition is given by $di_L/di_T = 0$ which from the first of the above equations yields

$$i_T^{(b-1)} = \frac{R}{bB}$$

Substituting this value of i_T in the expressions for i_L and V, the coordinates, V' and $i_{L(\max)}$, are obtained. In the neighborhood of V' the variation of i_L or of V_L, the potential across the load resistance, is least for any variation in V. Thus, if the potential V' is applied to the bridge, an approximately constant potential V_L appears across the load resistance even if the applied potential is subject to small variations. This supplies a means of obtaining an approximately constant potential from a fluctuating source. If the thyrite elements are replaced by lamps a similar condition can be obtained for fluctuations slow enough to permit the temperature of the filaments to alter. Such a lamp bridge is of use for alternating- as well as direct-current circuits.

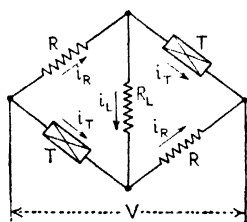



FIG. 5.4.—Thyrite bridge.

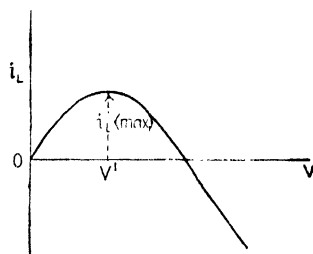



FIG. 5.5.—Thyrite-bridge characteristic.

5.4. Contingently Nonlinear Elements—Thermally Sensitive Resistances.—As these elements are characterized in general by the absence of any unique functional relationship between V and i it is not possible to reach simple general conclusions that are applicable to them as a class. If the external parameters upon which the effective resistance depends are held constant these elements may, of course, be treated in the same way as intrinsically nonlinear elements. Or if the values of these parameters are uniquely determined by i and V they may be again so treated. Elements for which the only additional parameter that need be considered in arriving at the characteristic is the temperature are known as *thermally sensitive resistances* and they serve as a good example of the behavior of contingently nonlinear elements in general. In this section the phenomtena presented by such elements in direct-current circuits will be considered. The steady-state alternating current applications are also readily tracable and will be considered in Sec. 5.8.

In the cases of many materials the specific resistance is given approximately by an exponential function of the reciprocal of the temperature

$$\rho = \rho_\infty e^{\frac{B}{T}} \qquad \text{or} \qquad \sigma = \sigma_\infty e^{-\frac{B}{T}} \tag{5.4}$$

These substances include metals, carbon, and the group of substances known as semiconductors. Metals have small specific resistances at room

temperature and semiconductors as a class have higher specific resistances at this temperature lying in the range from 10^{-3} to 10^7 ohm-meters. In certain types of semiconductors the current is carried in whole or in part by the motion of atomic ions. These are subject to polarization effects (Sec. 6.1) and the relation between i and V is not completely determined by the temperature. This type will be excluded from the following discussion which will be concerned only with those in which the current is electronic and the characteristic is determined by the temperature alone. Semiconductors are not as stable materials as metals and secular changes may occur with aging, but this complication is also neglected. Representative materials constituting semiconductors are given in Sec. 5.3. This list could be expanded to include many other oxides such as those of iron, manganese, nickel, and cobalt. The specific resistance at one temperature and the temperature coefficient are both very sensitive to certain types of impurities. The basic theory of such materials is of great interest and importance but lies beyond the scope of this treatment, and reference should be made to other texts.[1]

The coefficient B in Eq. (5.4) is frequently written as E/k where k is Boltzmann's constant. E is then of the dimensions of energy. For most metals E is very small, of the order of a few hundredths of an electron volt, and negative. For semiconductors E lies between 0.1 and 1.5 electron volts. For typical insulators E is several orders of magnitude greater than this. The coefficients ρ_∞ and σ_∞ in Eq. (5.4) represent obviously the values of ρ and σ for very great temperatures. It is generally more convenient to refer the resistance to its value at some fiducial temperature, say, T_0. As $R_0 = R_\infty^{\frac{B}{T_0}}$ the resistance may be written

$$R = R_0 e^{B\left(\frac{1}{T} - \frac{1}{T_0}\right)} \tag{5.5}$$

where R_0 is the resistance at $T = T_0$. The temperature coefficient of resistance, α, is defined as the fractional change in resistance per unit change in temperature or

$$\alpha_T = \frac{1}{R}\frac{dR}{dT} = -\frac{B}{T^2}$$

If B is very small as in the case of metals the exponential may be expanded in terms of $(T - T_0)$ and R may be written approximately as

$$R = R_0[1 + \alpha(T - T_0)] \tag{5.5'}$$

[1] WILSON, "Semiconductors and Metals," Cambridge University Press, London, 1939; MOTT and GURNEY, "Electronic Processes in Ionic Crystals," Oxford University Press, New York, 1940; SEITZ, "The Modern Theory of Solids," McGraw-Hill Book Company, Inc., New York, 1940.

α is approximately constant and independent of the temperature. Representative values of it for metals are given in Sec. 3.3.

Ordinary incandescent lamps represent circuit elements for which α is small but which can be operated over large enough temperature ranges to make the variation of resistance with temperature significant. The filament in a gas-filled lamp loses heat by conduction along the leads, conduction and convection through the gas, and by radiation so that an accurate thermal calculation for this case would be very difficult. On simplifying the problem by assuming a much smaller temperature range and considering only the loss of heat by straight conduction the general nature of the characteristic may be deduced from the known dependence of resistance on temperature. From Eq. (5.5′) the current through the filament is given by

$$i = \frac{V}{R_0[1 + \alpha(T - T_0)]}$$

Assuming that T_0 is also the temperature of the medium surrounding the filament the rate of loss of heat by conduction is proportional to $(T - T_0)$. The rate of generation of heat is iV so the condition of thermal equilibrium is

$$P = iV = \beta(T - T_0) \tag{5.6}$$

where β is a constant of proportionality depending on the filament and its surroundings. Eliminating the quantity $(T - T_0)$ from these two equations a quadratic equation is obtained for i. Choosing the appropriate root and expanding the radical by the binomial theorem in powers of the small quantity α, the current is found to be given by

$$i = \frac{1}{R_0}V - \frac{\alpha}{\beta R_0^2}V^3$$

to a first approximation. This is a cubic equation for i, and it is seen that if α is positive, the current increases less rapidly with the voltage than if the first term on the right were alone considered. This is the typical behavior of a metallic filament and such a characteristic is given by the solid curve of Fig. 5.6. In the case of carbon α is negative so the characteristic of a carbon filament resembles the lower dashed curve of this figure. Actually, of course, the dependence of resistance on temperature is more complicated than that which has been assumed and also the radiation from the hot filament plays an important role in cooling it.

The upper dashed curve of Fig. 5.6 represents the characteristic of a type of lamp used for reducing current fluctuations rather than for producing illumination. This is known as a *ballast lamp* and consists of an iron-alloy filament in an atmosphere of hydrogen. Its temperature coefficient of resistance is such that over a considerable

voltage range the current through it is approximately constant (30 to 60 volts in Fig. 5.6). Its principal use is in the reduction of line-voltage fluctuations in electrical circuits.

For semiconductors B is too large for Eq. (5.5′) to yield a satisfactory approximation, and Eq. (5.5) must be used. The equilibrium characteristic is obtained from Eqs. (5.5) and (5.6). In the cases of these elements the change in resistance is so large that they can be used as very satisfactory thermometers. In this application V and i are made so small that a negligible error is generally made in neglecting the generation of

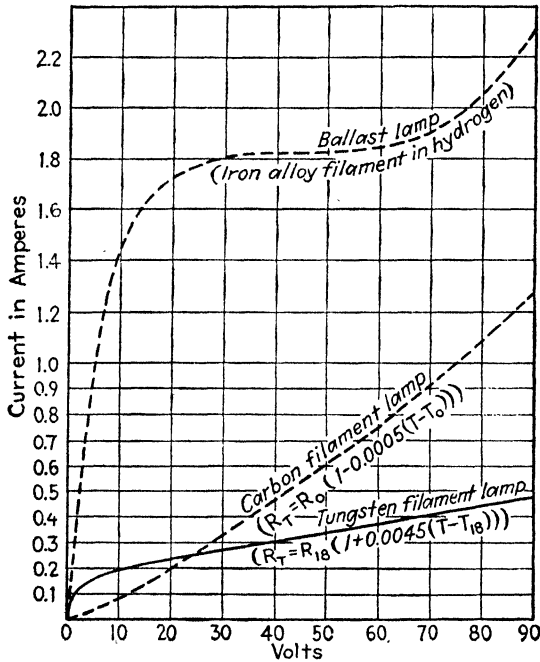



FIG. 5.6.—Typical lamp-filament characteristics.

power by the measuring circuit. They can also be used evidently as temperature-control devices where the magnitude of either the current or voltage when the other is held constant actuates control mechanism at some critical value. These temperature-sensitive elements, for which a representative characteristic is given by the curve in Fig. 5.1, are known as *thermistors,* and they are rapidly finding many applications in electric circuits.[1] As thermometers they can be used to measure temperatures to a precision of 5×10^{-4} °C. They can also be used as time delay switches, as can be seen by considering the change in resistance with time if the power expenditure in the thermistor is suddenly changed from

[1] BECKER, GREEN, and PEARSON, *Bell System Tech. J.*, **26**, 170 (1947).

one value to another or the external temperature, which may be taken as T_0, is changed to a new value T'_0. To represent such a change of conditions Eq. (5.6) must be extended to include the changing internal energy with temperature. If H is written as the effective specific heat, which will, of course, include the effects of leads and ambient material, the equation becomes

$$H\frac{dT}{dt} = P - \beta(T - T_0) \tag{5.6'}$$

Assuming for simplicity that negligible electric power is dissipated in the element ($P = 0$) and that $T = T'$ at $t = 0$, Eq. (5.6') can be integrated at once to give

$$(T - T_0) = (T' - T_0)e^{-\frac{t}{\tau}}$$

where $\tau = H/\beta$ is the relaxation time. The quantity T approaches T_0 exponentially, and the change in resistance can be determined from Eq. (5.5). An analogous expression is obtained if it is assumed that the temperature is suddenly increased from one value to another. In either case the lagging change in resistance can be made to operate a control circuit after a predetermined time interval.

5.5. Alternating Currents in Ohmic Circuits.—The most important uses of nonlinear elements are in connection with alternating rather than direct currents. Alternating currents are familiar as being the common means for the commercial supply of electric power; likewise at higher frequencies they form the basis of the communication industry. A treatment of the general theory of alternating currents must be postponed until the magnetic field has been introduced but alternating-current circuits containing only pure resistances may profitably be introduced at this point. The potential (difference) produced by an alternating-current generator and supplied to the distribution mains is ideally a sinusoidal function of the time. It may therefore be written

$$V = V_0 \sin \omega t$$

where V_0 is the maximum amplitude of the potential difference between the mains and ω is a constant known as the *angular velocity*. The plot of such a potential wave as a function of the time is given by the solid line of Fig. 5.7. It is the property of the sine function to reproduce itself exactly for every change in the argument of 2π. Thus, after the lapse of a time τ, V will return to the value it had at the beginning of the interval if $\omega t + 2\pi = \omega(t + \tau)$. This time interval τ, which from the above equation has the value $2\pi/\omega$, is known as the *period* of the potential wave. One full period is shown in Fig. 5.7; it is unnecessary to extend the time axis further, for all subsequent periods are exact repetitions

of this one. The reciprocal of τ is the number of periods which occur per unit time (second). It is known as the *frequency* and is represented by the symbol ν. Thus

$$\tau = \frac{1}{\nu} = \frac{2\pi}{\omega} \tag{5.7}$$

and the potential wave may be written alternatively as

$$V = V_0 \sin 2\pi\nu t$$

When an alternating potential is applied to the terminals of an ordinary ohmic resistance R, the current is at every instant proportional

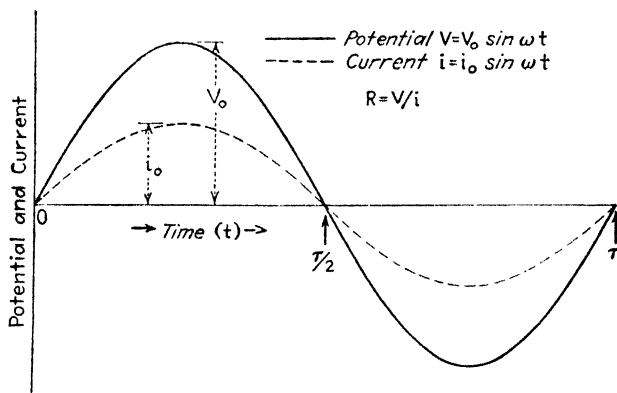



FIG. 5.7.—Alternating-current and -potential waves.

to the applied voltage, the constant of proportionality being $1/R$. Thus

$$i = \frac{V_0}{R} \sin 2\pi\nu t$$

or

$$i = i_0 \sin 2\pi\nu t$$

where i_0, which is written for V_0/R, is the maximum value of the current wave. The period is the same for both waves and the maximum and minimum points occur at the same time for both; this latter property is known as being *in phase*. The current i is represented by the dashed curve of Fig. 5.7. The instantaneous dissipation of power in the resistance is the product of V and i

$$P(t) = Vi = V_0 i_0 \sin^2 (2\pi\nu t)$$

Using the trigonometric identity $\sin^2 x = \frac{1}{2}(1 - \cos 2x)$

$$P(t) = \tfrac{1}{2}V_0 i_0(1 - \cos 4\pi\nu t)$$

This is seen to be a function of t with a frequency twice that of the potential or current waves; positive power is, of course, dissipated for

both senses of current flow. The average power dissipation is the integral of $P(t)$ over a period divided by the periodic time τ. Since the integral of the periodic term is found to be zero over this interval

$$P = \frac{1}{\tau}\int^{\tau} P(t)\, dt = \frac{1}{2}V_0 i_0 \tag{5.8}$$

This is the same as the dissipation of power by a constant potential and current V_e and i_e if $V_e = V_0/\sqrt{2}$ and $i_e = i_0/\sqrt{2}$. In consequence V_e and i_e are known as the *effective potential* and the *effective current* for the two waves.

If the current wave does not have its zeros and maxima at the same times as the potential wave the two are said to be out of phase. This is of frequent occurrence in general alternating-current circuits and its effect on the power consumption of the circuit is of great importance. If the dashed curve of Fig. 5.7 were slid to the right along the time axis an amount t', the figure would represent a current wave lagging the potential wave by a time t'. The equations for the waves would be

$$V = V_0 \sin 2\pi\nu t$$
$$i = i_0 \sin 2\pi\nu(t - t')$$

It is assumed that the current wave is not distorted but is merely shifted so that its characteristic features occur at a slightly later time. This, however, affects the power dissipated in the conductor. Forming the product Vi after expanding i in terms of the separate arguments

$$Vi = V_0 i_0(\sin^2 2\pi\nu t \cos 2\pi\nu t' - \sin 2\pi\nu t \cos 2\pi\nu t \sin 2\pi\nu t')$$

On integrating over a period and dividing by τ, the second term makes no contribution and the power dissipation is found to be

$$P = \tfrac{1}{2}V_0 i_0 \cos 2\pi\nu t'$$

or

$$P = V_e i_e \cos \varphi \tag{5.9}$$

Here φ, which is written for $2\pi\nu t'$, is the *phase lag*, and the factor $\cos \varphi$ which determines the power dissipation is called the *power factor*. This factor is of great importance in general alternating-current theory, but it is simply unity for the ohmic circuits that are here under discussion.

It is evident from the preceding discussion that alternating-current circuits made up of ohmic resistances are easily handled by ordinary algebraic methods. An arbitrary network is described at any instant by the system of linear equations developed for the general direct-current case [Eq. (4.4)]. Therefore these also apply in the case of alternating currents with the understanding that the $\mathcal{E}$'s and i's are sinusoidal functions of the time. Furthermore all the general circuit theorems developed in Sec. 4.3 are immediately applicable. However, nonlinear resistances

must generally be treated graphically and to illustrate the method, Fig. 5.8 represents the graphical analysis of an ohmic resistance in an alternating-current circuit. The straight line through the origin making an angle $\tan^{-1} R$ with the vertical current axis is the characteristic of the ohmic element. The lower extension of the current axis is considered as a time axis as well and the applied sinusoidal potential wave is drawn upon it. The right-hand extension of the potential axis also serves the purpose of a time axis and to the same scale. Choose a point corresponding to the time t on the potential curve. If a vertical line is drawn from this point on the curve, the ordinate of its intersection with the characteristic yields the current at time t. Hence the corresponding point on the current curve has this ordinate and the same abscissa t on the horizontal time scale. Following this procedure, the complete current curve may be drawn in on the horizontal time axis. Since the characteristic is straight, the current curve will be sinusoidal if the potential curve is; the procedure amounts to a simple linear transformation. However, the method is equally applicable if the characteristic is curved. This is the situation for nonlinear resistances and they will be found to give rise to distorted current waves. An example of this is seen in Fig. 5.11 which represents the distorted current wave resulting from the application of a sinusoidal potential wave to a thyrite element.

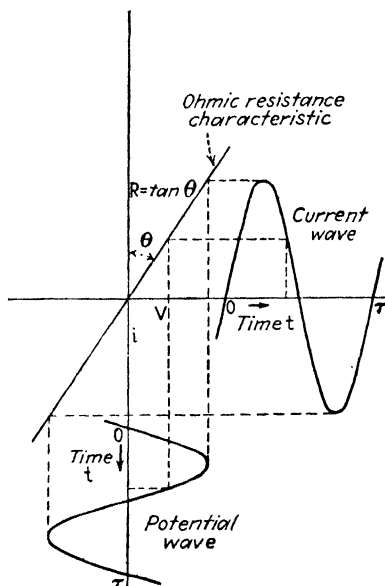



FIG. 5.8.—Graphical analysis of an ohmic resistance in an alternating-current circuit.

5.6. Alternating-current Circuits with Nonlinear Resistances.—Nonlinear resistances have many uses in alternating-current circuits. Lamp bridges, similar to the thyrite bridge of Fig. 5.4, may be used to reduce potential fluctuations which are slow enough to permit the temperatures of the filaments to change; elements such as thyrite may be used to produce odd harmonic distortion; and asymmetrical elements are used for rectification. At audio and radio frequencies these elements are used to mix two or more potential waves, for modulation, and for demodulation or detection. The general characteristics of nonlinear elements which enable them to perform these various functions will here be discussed both analytically and graphically.

In order to handle the problem analytically the appropriate characteristic of the element must be given in an analytic form. When a unique single-valued characteristic exists, i may be expressed by a Taylor's series in V. Utilizing Eq. (A.1) of Appendix A in which V is written for x, v_0 for x_0, i for $f(x)$, and i_0 for $f(x)_0$, the series for i in terms of V becomes

$$i = i_0 + \left(\frac{di}{dV}\right)_0 (V - V_0) + \frac{1}{2!}\left(\frac{d^2i}{dV^2}\right)_0 (V - V_0)^2 + \frac{1}{3!}\left(\frac{d^3i}{dV^3}\right)_0 (V - V_0)^3 + \cdots \quad (5.10)$$

The subscripts 0 indicate that the quantities are to be evaluated at $V = V_0$. If V is set equal to V_0 in Eq. (5.10), i becomes equal to i_0. The consumption of power for this value of V is $V_0 i_0$, and taking V_0 as positive, i_0 must also be positive unless the element contains a power source. (In this case $-V_0 i_0$ represents the static power output.) The ratio i_0/V_0 is the reciprocal of the apparent or static resistance at the point (V_0, i_0).

The coefficient of the first power of V, $(di/dV)_0$, is the slope of the characteristic at the point V_0 and is the reciprocal of the dynamic resistance. This must also be positive for an element that does not store power or contain a power source. For, if $(di/dV)_0$ is negative, the product $(V - V_0)(i - i_0)$ is negative to the first order for a small variation in V about V_0. Thus small oscillations would result in an output of power and by hypothesis there is no power source. If an element contains a power source and has a negative dynamic resistance, it may generate such oscillations. An element of this type is sometimes said to be "unstable." Thermistors represent elements with negative slopes over certain portions of the characteristic. Here the energy storage is in the form of heat. Other instances will be met in connection with vacuum tubes and gas discharges, and their properties will be discussed in connection with the generation of oscillations (Sec. 15.5). In this section it will be assumed that the rest of the circuit contains sufficient positive resistance to produce a positive slope for the over-all characteristic resulting in a stable circuit. A further point of interest about the coefficients in Eq. (5.10) is that those of even powers of V must be zero for symmetrical elements. Thyrite, lamp filaments, etc., are symmetrical elements for the characteristics are symmetrical in the sense that $i(V) = -i(-V)$ about $V = 0$. All powers of V may occur in the characteristic of an asymmetrical element. Rectifiers in general come in this category; in fact the coefficient of V^2 is a useful measure of the amount of rectification. Both odd and even powers of V will be retained in the following discussion.

Consider now that an alternating potential of the form $V = V_1 \cos \omega t$ is applied to such an element. Assuming expansion about the origin

($V_0 = i_0 = 0$), Eq. (5.10) becomes

$$i = i'V_1 \cos \omega t + \frac{i''}{2!}V_1^2 \cos^2 \omega t + \frac{i'''}{3!}V_1^3 \cos^3 \omega t + \frac{i''''}{4!}V_1^4 \cos^4 \omega t + \cdots$$

in which primes have been used to indicate the derivatives with respect to V; the evaluation is understood to be at $V = 0$. Using the following trigonometric identities

$$\begin{aligned} \cos^2 x &= \tfrac{1}{2}(\cos 2x + 1) \\ \cos^3 x &= \tfrac{1}{4}(\cos 3x + 3 \cos x) \\ \cos^4 x &= \tfrac{1}{8}(\cos 4x + 4 \cos 2x + 3) \end{aligned}$$

and collecting terms

$$\begin{aligned} i = {} & (\tfrac{1}{4}i''V_1^2 + \tfrac{1}{64}i''''V_1^4 + \cdots\cdots) \\ & + (i'V_1 + \tfrac{1}{8}i'''V_1^3 + \cdots\cdots) \cos \omega t \\ & + (\tfrac{1}{4}i''V_1^2 + \tfrac{1}{48}i''''V_1^4 + \cdots) \cos 2\omega t \\ & + (\tfrac{1}{24}i'''V_1^3 + \cdots\cdots\cdots\cdots) \cos 3\omega t \\ & + (\tfrac{1}{192}i''''V_1^4 + \cdots\cdots\cdots) \cos 4\omega t \end{aligned} \tag{5.11}$$

The current is seen to contain a constant term, a term of the original frequency, and terms with 2, 3, 4, etc., times this original or fundamental frequency. These latter terms are known as the second, third, fourth, etc., *harmonics*. The quantities in brackets give the amplitudes of these harmonics. The amplitudes of the constant term and the even harmonics are seen to involve only the coefficients of even powers of V in the original expansion and the coefficients of odd harmonics involve only coefficients of odd powers of V. Thus symmetrical elements give rise to only odd harmonics. Asymmetric elements in general give rise to a constant current in one direction or the other depending on the sign of the bracket and to all harmonics. It can also be seen that harmonics of order n or higher will appear only if the coefficients of V^n, V^{n+1}, etc., are appreciable.

Since Eq. (5.11) is the representation of i by means of a trigonometric series, it must be the same as the Fourier expansion of the characteristic in which i is considered as a function of $V_1 \cos \omega t$. Thus it could be derived equally well by the method of Appendix B. The harmonic terms which the series contains make no contribution to the power expenditure. This may be seen by forming the product Vi and integrating each term over a complete period of the fundamental. If A_n is the coefficient of the nth harmonic of the current curve, the contribution to the power made by this term is

$$P_n = \frac{1}{\tau}V_1A_n\int_0^\tau \cos \omega t \cos n\omega t \, dt$$

expanding the product of the cosines

$$P_n = \frac{1}{2\tau} V_1 A_n \int_0^\tau [\cos (n + 1)\omega t + \cos (n - 1)\omega t] \, dt$$

Both of these terms are zero when integrated over a complete period unless n is equal to unity, in which case the latter term contributes an amount τ. Thus $P_n = 0$ if n is not equal to unity and $P_1 = \frac{1}{2} V_1 A_1$, where A_1 is the coefficient of the $\cos \omega t$ term in Eq. (5.11). Examples of waves containing one harmonic in addition to the fundamental are shown in Fig. 5.9. The upper wave of that figure represents the type of distortion produced by a second harmonic with amplitude equal to the fundamental. The lower wave shows the distortion produced by a third harmonic component again of the amplitude of the fundamental. In the latter case if the resultant wave is shifted half a period along the time axis and rotated through 180° about this axis, it superposes upon the original wave. This is the type produced by a symmetrical nonlinear element.

Further points of interest emerge when a more complicated potential wave is applied to a nonlinear element. Assume the potential wave to be the sum of two sinusoidal components

$$V = V_1 \cos \omega_1 t + V_2 \cos \omega_2 t$$

To substitute this in Eq. (5.10), the quantities V^2, V^3, etc., must be calculated. Utilizing the trigonometric expressions for the products of cosines these become

$$V^2 = \tfrac{1}{2}(V_1^2 + V_2^2) + \tfrac{1}{2} V_1^2 \cos 2\omega_1 t + \tfrac{1}{2} V_2^2 \cos 2\omega_2 t + V_1 V_2 \cos (\omega_1 + \omega_2)t + V_1 V_2 \cos (\omega_1 - \omega_2)t$$

$$V^3 = \text{terms with the arguments: } \omega_1, \omega_2, 2\omega_1 \pm \omega_2, 2\omega_2 \pm \omega_1, 3\omega_1, 3\omega_2$$

and similarly for higher powers of V. The terms entering for V^n contain all possible arguments of the form $m_1\omega_1 \pm m_2\omega_2$, m_1 and m_2 being integers satisfying the equation $m_1 + m_2 = n - 2k$, where k takes on all values from zero to that which makes the sum of m_1 and m_2 either 1 or 0. Thus not only do all the harmonics of both frequencies occur but also all possible sum and difference frequencies. This thorough mixing of the two frequencies is very useful for many purposes. Consider the case of two alternating-current generators which have slightly different angular velocities so that $\omega_1 - \omega_2$ is equal to a small quantity $\Delta\omega$. If these two are placed in series with each other and a lamp (capable of being placed across a potential $V_1 + V_2$), the frequencies present in the line will be $\Delta\omega/2\pi$, $\omega/2\pi$, $(\omega \pm \Delta\omega)/2\pi$, and higher frequencies. Here ω_1 is taken as practically equal to ω_2 and is written ω. If it is an ordinary lamp and ω is a commercial frequency, the brightness of the filament will only

vary with the lowest frequency, $\Delta\omega/2\pi$. The lamp will thus record the electrical beats between the two generators. If the lamp is short-circuited at a dark phase of the filament, the generators will lock in with one another by their mutual interaction and they may subsequently be used in parallel.

A further interesting instance is that in which ω_2 corresponds to an audio frequency ν_a and ω_1 to a radio frequency ν_r. These two are widely different, having values of say 1,000 and 100,000 cycles, respectively. They are combined by a nonlinear element in a radio-transmitting

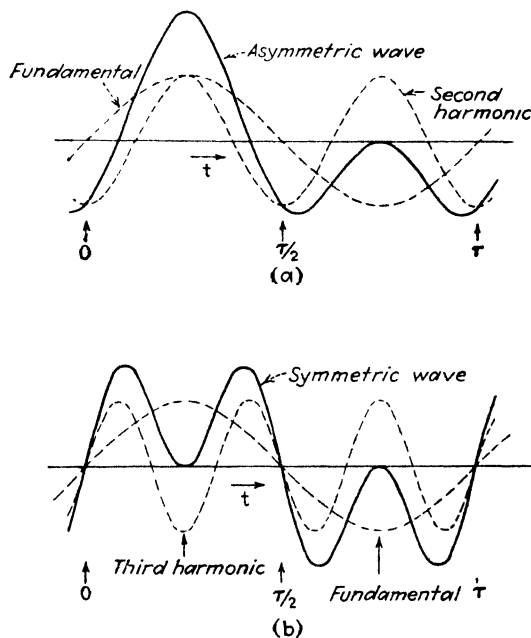



FIG. 5.9.—Examples of even- and odd-harmonic distortion.

station and the subsequent circulating currents contain the audio frequency ν_a, the radio frequencies ν_r, and $\nu_r \pm \nu_a$, as well as weaker components of the form $\nu_r \pm n\nu_a$, and higher radio frequencies of the order $2\nu_r$, $3\nu_r$, etc. Tuned circuits select from all these frequencies only those which are of the order of ν_r and they are radiated. The frequency ν_r is known as the *carrier* frequency and those of $\nu_r \pm \nu_a$ lying 1,000 cycles on either side are known as the *side bands*. The audio frequency cannot itself be directly radiated for reasons which will become apparent in the discussion of radiation, but it is intimately mixed with the radio frequency in the side bands. It is recovered at the receiving station by means of another nonlinear element. The original mixing process is known as *modulation* and the converse of it which takes place at

the receiving station is called *demodulation* or *detection*. The terms radiated may be written in various forms:

$$\begin{aligned} V_r &= V_1 \cos \omega_1 t + \frac{V_1 a}{2}[\cos (\omega_1 + \omega_2)t + \cos (\omega_1 - \omega_2)t] \\ &= V_1 \cos \omega_1 t + aV_1 \cos \omega_1 t \cos \omega_2 t \\ &= V_1 \cos \omega_1 t(1 + a \cos \omega_2 t) \end{aligned}$$

The last of these is generally the form in which the modulated wave is represented. The second term in the bracket may evidently be con-

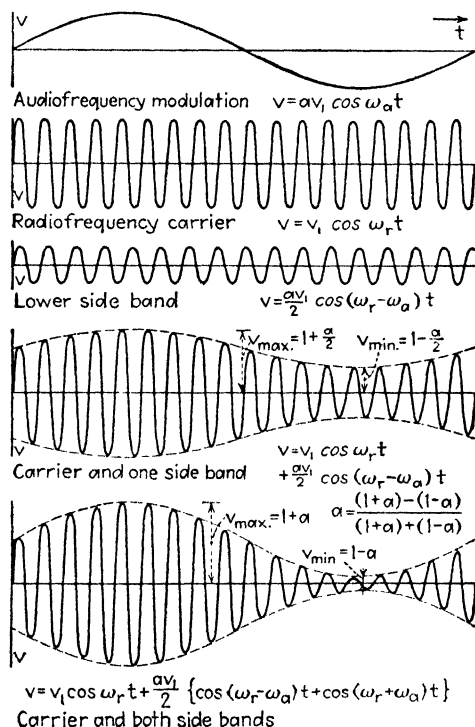



FIG. 5.10.—Sinusoidal waves associated with a modulated wave and the waves resulting from their composition.

sidered as imposing a variable amplitude of frequency $\omega_2/2\pi$ or ν_a on the radio-frequency oscillation. The quantity $100a$ is known as the *percentage modulation* of the wave. It is more convenient, however, to consider the first form as the potential wave applied to the nonlinear receiving element. Since it will give rise to all sum and difference frequencies of these three components, the receiving circuit will contain the frequencies corresponding to ω_2, $2\omega_2$, ω_1, $\omega_1 - \omega_2$, $\omega_1 + \omega_2$, etc. The radio frequencies are discarded and the coefficient of the harmonic $2\omega_2/2\pi$ should be small in comparison with that of the fundamental.

This desired audio frequency $\omega_2/2\pi$ is then amplified and delivered to the loud-speaker.[1]

5.7. Examples of Intrinsically Nonlinear Elements.—In the case of intrinsically nonlinear elements unique characteristics exist and in accordance with Eq. (5.10) these may be expanded either analytically or by graphical-numerical methods to yield the fundamental and harmonic components of the current resulting from a sinusoidal impressed emf. The general conclusions regarding the behavior of symmetrical and asymmetrical elements are, of course, applicable.

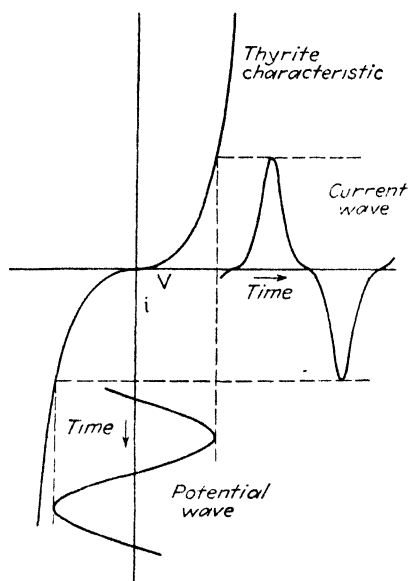



FIG. 5.11.—Sinusoidal potential wave applied to a thyrite element.

A good example of a symmetrical element is thyrite. A single thyrite element or a circuit containing both thyrite and ohmic resistances may be handled graphically, as shown in Fig. 5.11. The current curve is seen to contain only the odd type of harmonic distortion. For a single thyrite element the characteristic is of the form of Eq. (5.2), for thyrite in series with an ohmic resistance it is given by $Bi^{\frac{1}{a}} + Ri = V$. More complicated circuits may be analyzed in a similar way. The current wave through R_L of the thyrite bridge shown in Fig. 5.4, when a sinusoidal potential wave is applied to the input terminals, can be found from the bridge characteristic, Fig. 5.5. The harmonic content is found to vary widely with the amplitude of the potential wave. For a certain critical amplitude the fundamental is practically suppressed and only the third harmonic is in evidence.

The behavior in an alternating-current circuit of a single thyrite element, or any other circuit element obeying an equation of the form of Eq. (5.2) where a is positive, can also be treated analytically. It has been seen that the current curve will contain only odd harmonics of the fundamental period of the potential wave. Therefore the current

[1] The above refers to what is more precisely designated as *amplitude modulation* because the factor $(1 + a \cos \omega_2 t)$ essentially varies the amplitude V_1 at the angular frequency ω_2. Other types of modulation known as *frequency* and *phase modulation* vary $\omega_1 t$, the argument of the cosine function. There are some advantages in conveying intelligence in this alternative way, and for further details a text on radio engineering should be consulted.

wave may be written

$$i = i_1 \cos \omega t + i_3 \cos 3\omega t + i_5 \cos 5\omega t \cdots$$

From Appendix B, i_n is given by

$$i_n = \frac{1}{\pi}\int_{-\pi}^{\pi} i(\omega t) \cos n\omega t \, d(\omega t)$$

If V in Eq. (5.2) is replaced by the potential wave $V_0 \cos \omega t$, $i(\omega t)$ becomes

$$i(\omega t) = A V_0^a \cos^a \omega t$$

Substituting this in the above integral and reducing the range of integration from $(-\pi, \pi)$ to $(0, \pi/2)$, together with multiplication by 4 which the symmetry conditions permit, the coefficient of the fundamental is

$$i_1 = \frac{4A V_0^a}{\pi}\int_0^{\frac{\pi}{2}} \cos^{(a+1)} \omega t \, d(\omega t)$$

This definite integral is expressible as the ratio of two functions of the parameter a.

$$\int_0^{\frac{\pi}{2}} \cos^m x \, dx = \frac{\sqrt{\pi}}{2} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m+2}{2}\right)}$$

The function Γ is known as the gamma function. It obeys the relation $\Gamma(n + 1) = n\Gamma(n)$ which enables the function of a large argument to be reduced to that of smaller argument. If n is an integer, $\Gamma(n + 1)$ is evidently equal to factorial n. If n is not an integer, the argument of the function can be reduced by the above formula to a value between 1 and 2. Thus the function need be known between these two values only. A plot of the function through this range is given in Fig. 5.12. The function is of use for thermionic rectifiers as well as for thyrite. In terms of this function, i_1 becomes

$$i_1 = \frac{2A V_0^a}{\sqrt{\pi}} \frac{\Gamma\left(\frac{a+2}{2}\right)}{\Gamma\left(\frac{a+3}{2}\right)}$$

If a is equal to 3.5, the ratio of the gamma functions is found from the reduction formula and Fig. 5.12 to be 0.63. Furthermore, it may be shown that the following recurrence formula relates the coefficient i_n:

$$i_n = \frac{a - (n - 2)}{a + n} i_{(n-2)} \tag{5.12}$$

Calculating i_3, i_5, etc., from this relation, the current wave through a thyrite element for an impressed sinusoidal potential becomes

$$i = \frac{2.4\,V_0^{3.5}}{\sqrt{\pi}}(0.63 \cos \omega t + 0.23 \cos 3\omega t + 0.014 \cos 5\omega t \cdots)$$

This gives the magnitudes of the harmonic components directly; the relative harmonic content is seen to be independent of the amplitude of the potential wave. This method of analysis is not applicable to a circuit of thyrite and ohmic elements in series; such circuits must be treated graphically or by some method of successive approximations.

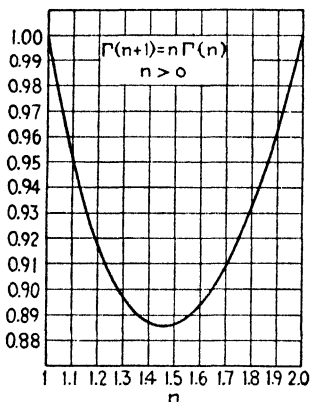



FIG. 5.12.—The gamma function, Γ (n).

An example of the asymmetrical element is the metal-semiconductor contact. To the extent that Eq. (5.3) is an adequate representation of the characteristic, it may be handled analytically in the obvious way of expanding i in powers of V and substituting $V = V_0 \cos \omega t$. This power series may be put in harmonic form by the series

$$\cos^n \omega t = \frac{n!}{2^{n-1}}\left[\frac{1}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!} + \frac{1}{\left(\frac{n-2}{2}\right)!\left(\frac{n+2}{2}\right)!}\cos 2\omega t + \cdots\right] \qquad (n \text{ even})$$

and the analogous series of terms of the form $\dfrac{1}{\left(\frac{n-m}{2}\right)!\left(\frac{n+m}{2}\right)}\cos m\omega t$

when n is odd. This then leads to

$$i(\omega t) = \sum A_m \cos m\omega t$$

where

$$A_m = \sum_{n=1}^{n=N} \frac{(bV_0)^n n!}{2^{n-1}\left(\frac{n-m}{2}\right)!\left(\frac{n+m}{2}\right)!}$$

here N is the exponent of the highest term retained in the power series expansion. A constant term ($m = 0$) and all higher harmonics to the Nth appear. The approximate nature of Eq. (5.3) does not warrant the retention of a large number of terms in the series, and more frequently this type of element is handled graphically.

Figure 5.13 shows a typical copper oxide rectifier characteristic at

about 20°C. It is seen to depart considerably from either Eq. (5.3) or what might be considered the ideal rectifier characteristic of Fig. 5.14. The graphical analysis of the current through an oxide element produced by a sinusoidal potential wave shows an appreciable current loop in the undesired sense. However, these elements are entirely satisfactory for

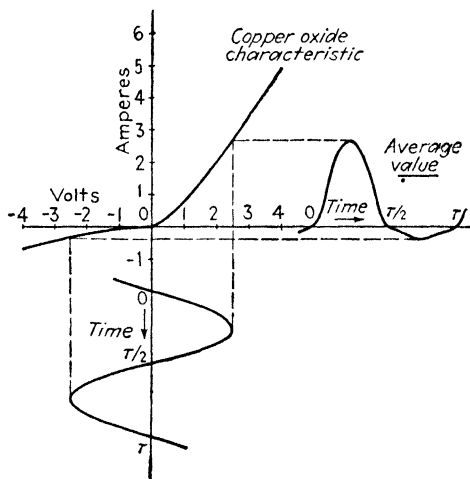



FIG. 5.13. Sinusoidal potential wave applied to a copper-oxide rectifier.

most purposes, as a small inverse component is generally permissible. The average value of the current passed in the positive direction for the wave of Fig. 5.13 is about 1.6 amp. The effectiveness of an element as a rectifier is measured by the ratio of the difference in the areas of the two current loops to the sum of these two areas. This evidently depends not only on the shape of the characteristic of the element but also on the amplitude of the applied potential wave. The current of Fig. 5.13 gives an effectiveness of about 75 per cent. The ideal rectifier characteristic of Fig. 5.14 evidently represents one that is 100 per cent effective for all potential waves. This idealized characteristic may be used to represent the behavior of actual oxide or thermionic rectifiers to a first approximation in the analysis of rectifier circuits.

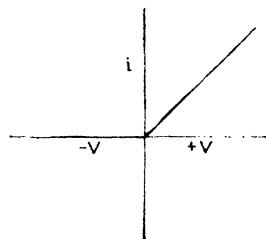



FIG. 5.14.—Idealized rectifier characteristic.

5.8. Examples of Contingently Nonlinear Elements.—The illustrations here will be drawn from thermally sensitive elements only, and it is convenient to divide these into two classes. The first is that in which the element is customarily used in the absence of any direct-current component. The ordinary incandescent lamp is a typical example. The second class is that in which the mean temperature is maintained by an

external source of heat, which may be a source of direct-current power, and an alternating emf. is superimposed. This is typical of the employment of thermistors in communication and control circuits. In both of the above cases the phenomena depend markedly upon the relative magnitudes of the period of the alternating emf. and the thermal relaxation time of the element under its conditions of use. If the fluctuation in

(a) Lamp filament.

(b) Thermistor.

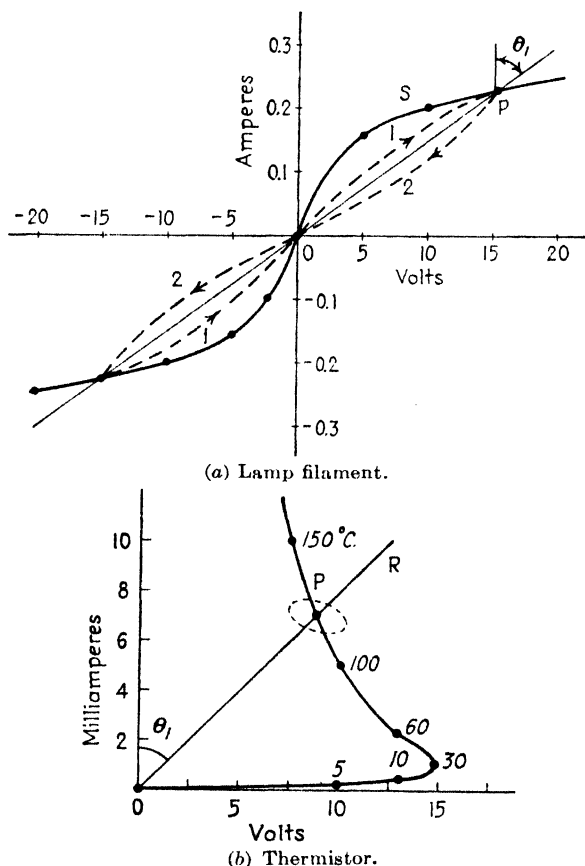



FIG. 5.15.—Contingent nonlinear elements in alternating current circuits.

current is very slow, the static characteristic will be followed approximately; and if it is so rapid that the temperature is effectively constant, the characteristic followed will be a straight line through the origin and the mean operating point. Intermediate values of the period of alternation lead to more complicated phenomena.

In Fig. 5.15*a* the curve *S* represents the characteristic of a metal lamp filament. If this is placed in an alternating-current circuit, the mean equilibrium temperature will be determined by the power expended in the lamp and the external conditions by Eq. (5.6). This establishes the

point P on the characteristic, which, in turn, determines the mean effective resistance. If the relaxation time is long compared with the period of current alternation, the straight line through P and the origin represents the effective characteristic that determines $i(V)$. As this is a straight line, the lamp is essentially a linear element under these conditions. If, however, the filament is very fine and the rate of heat loss from it large, τ becomes small, and in the limit of a slowly alternating current the curve S is traversed as a characteristic. This is curved but symmetrical, and hence only odd harmonics appear. For the intermediate case in which conditions are such that τ is of the same order as the period of alternation, the temperature fluctuates appreciably at double the frequency of the current, since the maximum rate of heating occurs twice per cycle. As the filament is colder when i and V are increasing and hotter when they are decreasing, the characteristic traversed is given qualitatively by the double-valued dashed curve of Fig. 5.15*a*. Cases of this type of characteristic are generally handled more conveniently graphically than analytically. However, this curve, in common with that of many other types of contingently nonlinear elements in alternating-current circuits, has a property which facilitates analytical discussion. As may be seen from the figure the curve is symmetrical about the origin in the sense that the branches 1 and 2 taken in sequence between the two extremes obey the condition $i_2(V) = -i_1(-V)$. This means that the expression for the alternating-current wave resulting from the application of a sinusoidal emf. has the property that its value at any instant is the negative of its value half a period before or after; *i.e.*, $i(\omega t) = -i(\omega t \pm \pi)$.* Taking a representative term in the Fourier expansion it is seen that the relation which must be fulfilled by each term is

$$\cos n\omega t = -\cos(n\omega t \pm n\pi)$$

This is true only if n is an odd integer as can be seen by expanding the right-hand term as products of sines and cosines. A similar condition shows that only odd n's can occur for sine terms as well, and thus the wave contains only odd harmonics. In general both sine and cosine terms appear in distinction to the case of a single-valued characteristic such as thyrite. This implies a shift in phase for the current components.

In Fig. 5.15*b* curve S represents the static characteristic of a thermistor. It will be assumed that it is connected in a direct-current circuit having adequate effective series resistance to give the composite characteristic a positive slope at the operating point P, which is determined by the power dissipated in the resistance and the rate of loss of heat to the

* This same condition is commonly encountered in dielectric and magnetic hysteresis.

surroundings. It will further be assumed that the other resistances in the circuit are very high or that they effectively present very high resistances to any alternating emf. impressed on the thermistor terminals in order that the thermistor alone need be considered in the behavior of the alternating components of the current. One of the most important applications of thermistors is in the measurement of small amounts of alternating-current power or the intensity of electromagnetic radiation. The most precise measurements of small powers are made with balanced bridges, one arm of which is the thermistor. The usual practice is to balance the bridge with direct current, measuring the amount of direct current in the nonlinear element at balance. The alternating-current power or electromagnetic radiation is then added to the thermistor, unbalancing the bridge. Balance is restored by removing some direct-current power from the thermistor, and the change in direct-current power is then equal in magnitude to the alternating-current power added. A sensitivity of several microwatts is quite common. The ultimate sensitivity depends on the stability of the bridge voltage supply and the compensation for variations in ambient temperature. A simple thermistor bridge circuit is shown in Fig. 5.16. One arm of the Wheatstone bridge consists of a thermistor, and the other three arms are temperature independent and equal to the resistance of the thermistor at the operating point. A variable resistance R' is placed in series with a resistanceless battery of emf. $\mathcal{E}$. The current supplied by the battery at balance is $i = \dfrac{\mathcal{E}}{R' + R}$, and the power dissipated in the thermistor is $R\left(\dfrac{i}{2}\right)^2 = \dfrac{R\mathcal{E}^2}{4(R' + R)^2}$. If a small amount of power P_0 supplied from an external source is dissipated in the thermistor arm, then R' must be increased by an amount $\Delta R'$ in order that the total power dissipated in the thermistor remain constant. Thus

FIG. 5.16.—Measurement of high frequency power with a thermistor.

$$P_0 = \frac{R\mathcal{E}^2\,\Delta R'}{2(R' + R)^3}$$

Bridges of this type with a sensitivity of 10^{-6} watt per galvanometer division can be readily constructed.

If the alternating emf. applied to the thermistor is neither so rapid that the straight characteristic from P through the origin is followed nor so slow that the curve S is followed, the situation requires a more detailed analysis. Assume that the mean resistance characterized by the point P is R_p. Then from Eq. (5.5)

$$R = R_p e^{B\left(\frac{1}{T} - \frac{1}{T_p}\right)} = R_p'(1 + \alpha_p(T - T_p))$$

where α_p is the temperature coefficient at P. Then writing x for $(T - T_p)$ where x is assumed small, Eq. (5.6′) is

$$\tau \frac{dx}{dt} + x = \frac{P}{\beta} - (T_p - T_0)$$

where T_0 is the ambient temperature and τ, the relaxation time, is H/β. If it is assumed that the emf. applied to the element is $V = V_0(1 + V' \cos \omega t)$, the equation will be satisfied if the current is given by $i = i_0[1 + i' \cos (\omega t + \psi)]$, where V' and i' are considerably less than unity. Inserting Vi for P and V/i for R in $\alpha x = (R - R_p)/R_p$ the terms of zero order in the equation yield

$$P_0 = V_0 i_0 = \beta(T_p - T_0)$$

The first-order terms in V' and i' yield equations that can be solved for ψ and i':

$$\tan \psi = \frac{2\omega\tau C}{(\omega\tau)^2 + (1 - C^2)}$$

$$i' = \left[\frac{(1 - C)^2 + (\omega\tau)^2}{(1 + C)^2 + (\omega\tau)^2}\right]^{1/2} V'$$

where $C = \alpha_p P_0/\beta$. Thus the phase difference between the alternating component of i and V and the amplitude of the alternating component of i depend on ω, τ, and the parameter C. If $\omega\tau$ is either very large or very small, ψ approaches zero and the voltage and current waves are in phase in accordance with the previous qualitative conclusions. The amplitudes of i' differ in the two cases, being V' in the former and $V'(1 - C)/(1 + C)$ in the latter. The factor $(1 - C)/(1 + C)$ may be shown to be the ratio of the static to dynamic resistance at P, corresponding to motion along the static characteristic for very low frequencies. For intermediate frequencies it is evident that a finite phase difference exists between i' and V', which implies that the actual characteristic traversed is a small closed curve about the point P.

The nature of this curve and the sense in which it is traversed depend on the sign and magnitude of the parameter C. By Eq. (5.6) $C = \alpha_p(T_p - T_0)$. For metals α_p is very small and positive so ψ is positive but very small for all values of ω and i' is approximately equal to V'.

For semiconductors such as thermistors α_p is larger and negative of the order of -0.03 per degree centigrade. Thus if $T_b - T_0$ exceeds about 30°C., $\tan \psi$ is negative for small values of $\omega\tau$ and becomes infinite for $\omega\tau = (C^2 - 1)$; for increasing values of ω it approaches zero through positive values,[1] both numerator and denominator being negative. The negative value of the denominator is associated with the negative slope of the thermistor characteristic and corresponds to the fact that the thermistor under these conditions is not really a passive element but is a device that can draw on the direct-current power source to sustain small amplitude oscillations. The current amplitude, for values of $C < -1$, varies smoothly from $(1 - C)/(1 + C)V'$ to V' for values of $\omega\tau$ from zero to infinity.

5.9. Rectifier Circuits.[2]—Oxide rectifiers are useful for relatively low alternating-current voltages. In the higher range, roughly from 100 to 25,000 volts, the high-vacuum thermionic rectifier or gas discharge is used. For still higher voltages in the typical X-ray range only high vacuum tubes can be employed successfully. These three different types of rectifier all have different characteristic curves. The high-vacuum type obeys a law of the form $i = 0$ for $V < 0$ and $i = AV^{3/2}$ for $V > 0$. The typical thermionic mercury-vapor rectifier passes no current in the negative sense and operates at a constant potential drop of from 10 to 15 volts when conducting. Thus in this case the current is limited only by the resistance load in series with the rectifier. This imposes certain conditions on the type of load circuit that can be employed. If a large condenser is placed directly in series with such a discharge, in an alternating-current circuit, it presents very little resistance to the sudden surge of current which takes place when the rectifier breaks down and becomes conducting. As a consequence very large instantaneous currents will be drawn which may injure the cathode of the rectifier. For this reason and to suppress oscillations that may result from instability a series resistance (or reactance) is generally employed with this type of element. Owing to the constant potential drop in such an element the fraction of the total power dissipated in it is independent of the series load for any applied voltage. For the high-vacuum rectifier, however, the drop across it increases with increasing current and the fraction of the power dissipated in it increases as the

[1] In the discussion of general alternating-current circuits it will be seen that a negative ψ corresponds to an inductive reactance and a positive ψ to a capacitative reactance. The changing of the denominator from negative to positive values for increasing ω corresponds to changing from negative to positive values of effective resistance at $\omega\tau = (C^2 - 1)$. At this point the resistive reactance of the thermistor is zero.

[2] JOLLEY, "Alternating Current Rectifiers," John Wiley & Sons, Inc., New York, 1925.

resistance of a series load is decreased. For both types of rectifier the efficiency increases with increasing applied voltage.

If the resistance represented by the line in the upper right-hand quadrant of Fig. 5.14 is the total resistance in the circuit, the figure represents very accurately the over-all characteristic of a vapor-rectifier circuit. The current axis, however, must be shifted to the left by an amount equal to the drop in the rectifier. The same figure just as it stands may be used to represent approximately the behavior of the high-vacuum or oxide rectifier circuit. The approximation becomes better as the series resistance is increased for then the nonlinear characteristic of the rectifier element itself becomes of less importance. In the following brief discussion of various rectifier circuits this type of characteristic is assumed.

Figure 5.17 represents the simplest type of rectifier circuit and the current wave passing through it under the influence of an applied sinusoidal potential. The positive-current loops are sinusoidal and the negative ones are entirely suppressed. Such a circuit is known as a half-wave rectifier. The current is obviously a pulsating one with a direct-current component and a series of alternating-current harmonics. For the general circuit calculations that can be carried out with the aid of the alternating-current theory of later chapters it is necessary to know the amplitudes of the various harmonics. The Fourier analysis of such a wave (Appendix *B*) is relatively simple. Choosing the zero of time at a current maximum, $i(t) = i(-t)$. This is characteristic of the cosine function and hence i may be represented as a cosine series. Writing x for ωt

$$i = \frac{a_0}{2} + a_1 \cos x + a_2 \cos 2x + a_3 \cos 3x + a_4 \cos 4x + \cdots$$

From the curve of Fig. 5.17

$$i = 0 \text{ for } -\pi < x < -\frac{\pi}{2} \text{ and } \frac{\pi}{2} < x < \pi \quad \text{and} \quad i = I \cos x \text{ for } -\frac{\pi}{2} < x < \frac{\pi}{2}$$

Hence

$$\begin{aligned} a_n &= \frac{I}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos x \cos nx \, dx \\ &= \frac{I}{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} [\cos (n-1)x + \cos (n+1)x] \, dx \\ &= \frac{I}{2\pi}\left[\frac{\sin (n-1)x}{n-1} + \frac{\sin (n+1)x}{n+1}\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \end{aligned}$$

If n is odd, this expression vanishes unless $n = 1$, in which case the first term becomes $I/2$. If n is even, the expression is seen to reduce to

$$a_n = -\frac{2I}{\pi}\frac{1}{n^2 - 1}\cos\left(\frac{n\pi}{2}\right).$$

Thus

$$i = \frac{2I}{\pi}\left(\frac{1}{2} + \frac{\pi}{4}\cos\omega t + \frac{1}{3}\cos 2\omega t - \frac{1}{15}\cos 4\omega t + \frac{1}{35}\cos 6\omega t \cdots\right) \tag{5.13}$$

I may also be written as V_0/R, where V_0 is the peak alternating-current potential applied to the circuit and R is the effective resistance during the conducting half cycle. The amplitude of the first harmonic is $\pi/2$ times the amplitude of the constant term and the second harmonic has $\frac{2}{3}$ the amplitude of the constant term.

FIG. 5.17.—Half-wave rectification.

Figure 5.18 (a and b) shows two circuit arrangements for utilizing both halves of the potential wave. If the mid-point of the alternating-current circuit is available as in the case of a center-tapped transformer or with a center-tapped resistance across the line, the simple circuit shown at a may be used. In this case the peak potential applied to the load is only half the peak potential between the alternating-current lines but the potential applied to the rectifier in the reverse direction is equal to

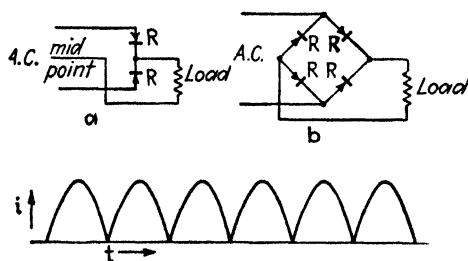



FIG. 5.18.—Full-wave rectification.

the peak potential. At b four rectifier elements are shown in a bridge circuit. This accomplishes the same purpose without the use of a mid-point. The potential across the rectifier system and load is the same as that across the line; the effective rectifier resistance in series with the load is twice that of a single element. The current wave produced by such circuits is shown in the same figure. This is evidently equal to the sum of the expression for the current represented by Fig. 5.17 plus a

similar expression displaced along the time scale half a period. On adding to Eq. (5.13) a similar expression with ωt replaced by $(\omega t + \pi)$, we obtain

$$i = \frac{4I}{\pi}\left(\frac{1}{2} + \frac{1}{3}\cos 2\omega t - \frac{1}{15}\cos 4\omega t + \frac{1}{35}\cos 6\omega t \cdots\right)$$

This represents the current output of an ideal full-wave rectifier. The fundamental or first harmonic is absent; otherwise the current is twice that from a half-wave rectifier.

The use of condensers in a rectifier circuit permits the output voltage to be increased to any integral multiple of the alternating-current voltage. Figure 5.19 shows a circuit for supplying a unidirectional voltage equal to twice the peak alternating potential. Neglecting the load, it is seen that each condenser is charged to the peak value of the alternating-current wave once each cycle. The condensers are charged in such a sense that their potentials add together and the total voltage across them is thus twice the peak value of the alternating potential. The circuit constitutes a full-wave rectifier with a current wave similar to that in Fig. 5.18. An analysis of the potential appearing across the load required the use of the general alternating-current theory of resistance-capacity circuits. However, the approximate ratio of the magnitude of the second harmonic component to the direct-current component can be found from elementary considerations. Each condenser is charged once a cycle to a potential $V/2$, where V is the potential appearing across the load. Then the charge flows out of this condenser through the load resistance R for a time approximately equal to the period τ of the alternating-current cycle. This current may be written approximately

FIG. 5.19.—Rectifying and potential-doubling circuit.

$$i = \frac{V}{R} = \frac{\Delta q}{\tau} = \frac{C\Delta V}{2\tau}$$

or

$$\frac{\Delta V}{V} = \frac{2\tau}{RC}$$

Thus the percentage voltage fluctuation across the load is seen to be of the order of the ratio of twice the alternating-current period to the time constant of the circuit.

Very commonly a so-called three-phase alternating-current distributing system is used. The potential of each of three wires fluctuates sinusoidally with respect to the ground or any neutral point. These potential waves are not in phase but are equally spaced in time so that the

maxima of any two of these waves are separated by a time $\tau/3$. This corresponds to a phase lag of $2\pi/3$, or 120°. By the use of three elements such a system of waves may be rectified partially, or by means of six, full-wave rectification can be obtained. Figure 5.20 shows a circuit for full-wave rectification and also the components making up the final current output. The sum of these components is seen to recur periodically after each interval of time of length $\tau/6$. This means of course that the lowest harmonic appearing is that which has a period of $\tau/6$ or the 6th harmonic of the original alternating-current wave. The series thus

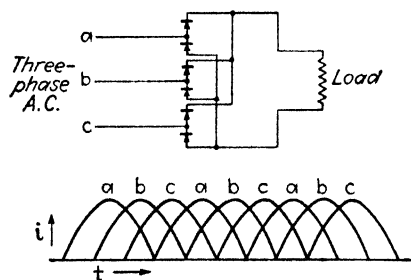



FIG. 5.20.—Three-phase full-wave-rectification.

contains a constant term and the 6th, 12th, 18th, etc., harmonics. Adding together the three waves from $-\frac{\pi}{6}$ to $+\frac{\pi}{6}$ about a point of symmetry

$$\begin{aligned} i &= I \cos \omega t + I \cos \left(\omega t + \frac{\pi}{3}\right) + I \cos \left(\omega t - \frac{\pi}{3}\right) \\ &= 2\, I \cos \omega t \end{aligned}$$

The symmetry is such that only cosine terms appear and the coefficients a_n are given by the following integral:

$$\begin{aligned} a_n &= \frac{2 \times 12}{\pi} I \int_0^{\frac{\pi}{6}} \cos \omega t \cos n\omega t \, d(\omega t) \\ &= \frac{I}{\pi} \frac{12}{(n^2 - 1)} \text{ with alternate } + \text{ and } - \text{ signs} \end{aligned}$$

Thus the current may be written

$$i = \frac{12}{\pi} I \left(\frac{1}{2} + \frac{1}{35} \cos 6\omega t - \frac{1}{143} \cos 12\omega t + \cdots \right)$$

Here I is the amplitude of an individual current wave. The amplitude of the first periodic term to occur in the series which is the 6th harmonic is only about 6 per cent of the constant term. Thus the periodic terms in the output of a full-wave three-phase rectifier are small and for many purposes they may be entirely neglected.

Problems

1. A mercury-vapor rectifier operating with a 15-volt drop is in series with a resistance of 100 ohms. Calculate the voltage drop across each element, the total circuit voltage, and the ratio of the power lost in the rectifier to that supplied to the circuit for the following direct currents: (*a*) 1 amp.; (*b*) 0.064 amp.

2. A high-vacuum thermionic rectifier with the characteristic $i = 10^{-2}V^{3/2}$ (in amperes and volts) is in series with a resistance of 100 ohms. Calculate the voltage drop across each element, the total circuit voltage, and the ratio of the power lost in the rectifier to that supplied to the circuit for the same two currents as in the preceding problem.

3. A bridge circuit is constructed of elements of thyrite for which $A = 3.1 \times 10^{-8}$ (in units of volts and amperes) and resistances of 100 ohms. Calculate the applied voltage for which there is no current through the load resistance.

4. Using the constants of the preceding problem and a load resistance of 1,000 ohms, calculate: the current through the load; the voltage across it; and the potential applied to the bridge for the condition of minimum output variation.

5. An alternating voltage given by $V = 3 \sin \omega t$ is applied to the rectifier bridge of Fig. 5.18*b*. If the effective resistance of a rectifier element is 1 ohm and that of the load is 3 ohms, calculate the maximum voltage drop across the load and the direct-current component of the current through it.

6. Find both analytically and graphically the current wave through a rectifying element with the characteristic: $i = -0.1V$ for $-10 < V < 0$ and $i = V^2$ for $0 < V < 10$ with an applied voltage wave $V = 5 \sin \omega t$. Find the charge passed in each direction per cycle.

7. Calculate the amplitudes of the constant term and the harmonics for the current wave when a potential $V_0 \sin \omega t$ is applied to an element for which $i = aV(V + 2b)$ (where V_0 is less than the absolute magnitude of b). Plot the characteristic and analyze the wave graphically as well.

8. Two sinusoidal voltage waves $6 \sin 2\pi\nu_1 t$ and $3 \sin 2\pi\nu_2 t$ (where, for instance, ν_1 is 100,000 and ν_2 is 1,000 cycles per second) are applied to a nonlinear element with a characteristic of the form $i = V(V + 10)$. Calculate the amplitudes of the current components. Note which of these is the carrier, which are the side bands, etc.

9. An 80 per-cent-modulated voltage wave $V = 5 \cos 2\pi\nu_1 t(1 + 0.8 \cos 2\pi\nu_2 t)$ is applied to the nonlinear element of the preceding problem. Calculate the amplitudes of the constant term and the various harmonic components.

10. Using the cubic approximation for the characteristic of a filament (Sec. 5.4) calculate the amplitudes of the current harmonics for a sinusoidal potential wave of very long period.

11. Plot the over-all characteristic of a circuit containing a thyrite element for which $A = 3.1 \times 10^{-8}$ (units of amperes and volts) and a series resistance of 100 ohms from 0 to 200 volts. Use it to analyze graphically the current produced by the voltage $150 \sin \omega t$.

12. If a potential wave $2V_0 \sin \omega t$ is applied to the rectifier circuit of Fig. 5.18*a* and the rectifiers are mercury-vapor tubes with a drop of V' when conducting, derive the expressions for the Fourier coefficients of the current.

13. Draw a circuit for a half-wave three-phase rectifier and show that the output current is given by

$$i = \frac{6I}{\pi}\left(\frac{1}{2} + \frac{1}{35} \cos 6\omega t - \frac{1}{143} \cos 12\omega t + \cdots \right)$$

where I is the amplitude of a single current wave.

14. A voltage wave $V = 150 \sin \omega t$ is applied to a circuit made up of a resistance of 47.7 ohms and a mercury-vapor rectifier operating at a potential drop of 15 volts when conducting. Find the current wave.

15. A potential wave of the form $V = V_0 \cos \omega t$ is applied to the terminals of a thermionic rectifier for which $i = AV^{3/2}$ for $V > 0$ and $i = 0$ for $V < 0$. Show by means of the gamma function that i can be written

$$i = \frac{AV_0^{3/2}}{\sqrt{\pi}}(0.493 + 0.810 \cos \omega t + 0.423 \cos 2\omega t + 0.090 \cos 3\omega t - 0.038 \cos 4\omega t \cdots)$$

[the recursion formula Eq. (5.12) applies].

16. Plot the ratio of the dynamic to the static resistance as a function of V using the characteristic of Prob. 7.

17. Assuming a parabolic characteristic, as in Prob. 7, and a sinusoidal applied potential, show that the ratios of both the unidirectional term and the amplitude of the second harmonic to that of the first harmonic are given by

$$\frac{1}{2}\frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

where I_{max} and I_{min} are the absolute values of the positive and negative current peaks, respectively.

18. Assuming a cubic characteristic, $i = aV + bV^3$, show that the ratio of the amplitude of the third harmonic to that of the fundamental for a sinusoidal applied emf. is given by

$$\frac{1}{2}\frac{I_{max} - 2I_x}{I_{max} + I_x}$$

where I_{max} is the maximum value of the current and I_x is the magnitude of the current wave $\frac{1}{12}$ of a period from a current zero.

19. A condenser of capacity C which is charged to a potential V_0 at a time $t = 0$ discharges through an element with the characteristic $i = AV^n$. Show that the potential at any later time t is given by

$$V = \left[\frac{(n-1)At}{C} + V_0^{-(n-1)}\right]^{-\frac{1}{(n-1)}}$$

20. Show that the conduction characteristic $i = AV^a$ can be formally accounted for if the conduction electrons are retarded by a force $\left(\frac{n}{A'}ve^{(a+1)}\right)^{\frac{1}{a}}$, where e is the electronic charge, v the drift velocity, and n is the number of conduction electrons per unit volume. A' is A multiplied by the length of the specimen to the power a and divided by the area of cross section.

21. A single side band of the form

$$V_s = V_1 \cos(\omega_r - \omega_a)t$$

is added linearly to a potential wave of the carrier frequency

$$V_c = V_2 \cos \omega_r t$$

Show that the combination can be written

$$V_s + V_c = (V_1^2 + V_2^2 + 2V_1V_2 \cos \omega_a t)^{1/2} \cos(\omega_r + \phi)$$

where

$$\phi = \tan^{-1} \frac{\sin \omega_a t}{\cos \omega_a t + \dfrac{V_2}{V_1}}$$

which represents a 100 per cent modulated wave if $V_1 = V_2$.

22. A potential wave of the form $V = V_1 \cos \omega_r t$ is applied to an antenna which may be thought of as a pure resistance. The same wave is then modulated a fraction a and applied to the same antenna. Show that the power P consumed at the carrier frequency is the same in both cases and that in the second case an additional amount $(a/2)^2P$ is consumed at each of the side-band frequencies. Assuming that the carrier frequency is available at the receiver (hence only one side band need be radiated), show that $2 + (2/a)^2$ as much useful power can be radiated by the transmitter if all components but one side band are suppressed.

23. The characteristic of a copper oxide rectifier for small voltages is given by Eq. (5.3). Show that if a small sinusoidal voltage is applied, the rectified current is $b/2$ times the power consumed. Find the ratio of the amplitude of the current of the fundamental frequency to that of the second harmonic.

24. Assuming that four rectifier elements having characteristics given by Eq. (5.3) are arranged in a bridge as shown in Fig. 5.18 with a load resistance R, show that the current through the load is given in terms of the applied voltage by the following transcendental equation

$$\cosh \frac{\alpha V}{2} = \frac{1}{2A} i e^{\frac{bRi}{2}}$$

Determine the effective input resistance of the bridge, and show how the equations can be solved graphically.

25. Using Eqs. (5.5) and (5.6), show that the extreme values of V for a thermistor occur at the temperatures

$$T_m = \frac{B}{2}\left(1 \pm \sqrt{1 - \frac{4T_0}{B}}\right)$$

where the minus sign corresponds to a maximum and the plus sign to a minimum. Derive the extreme values for the resistance, current, and potential difference.

26. Plot the thermistor characteristic of Fig. 5.1 on a linear scale, and determine graphically the value of the resistance that, when placed in series with it, will yield the composite characteristic having an extended region representing an approximately constant potential difference across the two elements.

27. For a thermally sensitive resistance show that the ratio of the dynamic resistance $R_p' = (dV/di)_p$ to the static resistance $R_p = (V/i)_p$ at the point p is given by

$$\frac{R_p'}{R_p} = \frac{1 + (\alpha_p P_0/\beta)}{1 - (\alpha_p P_0/\beta)}$$

where α_p is the temperature coefficient of resistance, P_0 is the direct-current power being supplied, and $\beta = P_0/(T_p - T_0)$, where T_0 is the ambient temperature.

CHAPTER VI

CHEMICAL, THERMAL, AND PHOTOELECTRIC EFFECTS

6.1. Conduction of Electricity in Liquids.—Liquids may be roughly divided into three classes on the basis of their conductivities. It was seen in Sec. 3.2 that certain ones have very low conductivities of the order of 10^{-11} mho per meter. The paraffin oils and certain aromatic liquids such as xylol are representative of this class. They are widely used as insulators and dielectrics. Then there is the class of pure liquids represented by the alcohols and water, which have conductivities of the order of 10^7 times as great. These are in general unsatisfactory as insulators or as conductors. The third class is represented by a solution of an acid, base, or salt in water. These solutions have conductivities of the order of 10^5 times that of pure water. For many purposes they may be considered as good conductors though their conductivities are less by a factor of about 10^{-5} than the typical metallic conductors. These solutions of large conductivity are known as *electrolytic solutions* and the solute (dissolved substance) is called an *electrolyte*.

Electrolytic conduction follows approximately the ohmic law. However, the current is not carried by electrons as in the case of metallic conduction but by the more massive positive and negative ions that are present in great abundance in such solutions. This is demonstrated by the deposition of the components of the dissolved substance at the electrodes when a current flows between them. Consider a solution of a copper salt in which are inserted two copper electrodes. If an electromotive force is applied between them, a current is observed to pass from one to the other through the solution. This current flow is accompanied by a wasting of the *anode* (positive electrode) and a deposition of copper on the *cathode* (negative electrode). The sum of the masses of the two electrodes remains constant. The amount of copper transferred from one electrode to the other is found to be proportional both to the current strength and to the time of flow, *i.e.*, it is proportional to the total charge which has passed through the solution. If a similar experiment were performed with some other element, say a silver salt solution and silver electrodes, the same phenomenon would be observed. If these two cells were placed in series in an electric circuit, it would be found that the ratio of the mass of the copper deposited to that of the silver was the same as the ratio of the chemical equivalents of the two

substances. The chemical equivalent is the atomic weight divided by the charge per ion, or valence.

A systematic study of the phenomena accompanying electrolytic conduction was first made by Faraday and on the basis of his observations he formulated the following laws:

The mass of a substance liberated at an electrode by the passage of an electric current is proportional to the total charge that has passed.

The mass of a substance liberated at an electrode is proportional to the chemical equivalent of the substance.

Writing m for the mass in grams, M for the atomic or molecular weight, and v for the valence, these two laws may be expressed by the equation

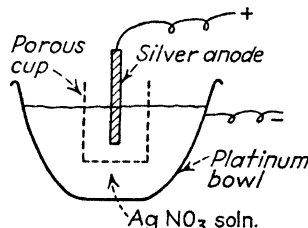
$$m = \frac{Mit}{vF}$$

where F is the constant of proportionality that must be introduced. This constant is known as the *faraday*. If N is the number of atoms per mole (one mole is equal to M gm. of the substance) and n is the number of atoms transferred from one electrode to the other, then $M/m = N/n$. Each ion carries with it a charge ve, where e is the electronic charge. Therefore the total charge passed it is nve. Making these substitutions in the above equation, it is found that

FIG. 6.1.—Silver voltameter.

$$F = Ne$$

Thus F is the product of these two fundamental atomic constants. This relation provides the most accurate method of determining N in terms of the experimental values of F and e.

The faraday is by definition the product of the chemical equivalent and the total charge passed divided by the mass deposited. The experimental cell for determining the mass in terms of the current and time is known as a *voltameter*. Two of these are standard, one the copper voltameter and the other the silver voltameter. The copper voltameter consists of two copper electrodes immersed in a 25 per cent solution of $CuSO_4$ acidulated with 1 per cent H_2SO_4. The standard current density is 20 ma. per square centimeter of cathode surface, and it is the increase in mass of the cathode that is measured. It is found that 0.0003294 gm. of copper is deposited for each coulomb of electricity that passes. A standard form of the silver voltameter is shown in Fig. 6.1. A platinum bowl forms the cathode and it contains a 15 or 20 per cent solution of $AgNO_3$. A silver anode immersed in this solution is surrounded by a porous cup to prevent any particles which may become loosened from falling on the cathode and being occluded in the electrically deposited layer. For details of the precautions which must be taken to achieve the highest accuracy reference should be made to a more complete account.[1] As a high accuracy can be achieved in this type of experi-

[1] BIRGE, Reports on Progress in Physics, *Phys. Soc. London*, **8**, 112 (1941).

ment, it was used prior to 1948 to provide the legal or international definition of the ampere. An ampere was defined in that system to be the current which when flowing through a properly designed silver voltameter deposits silver upon the cathode at the rate of 0.00111800 gm. per second. This definition was chosen in order to make the international ampere equal as nearly as possible to the absolute practical ampere which is now the unit of current. The most careful experiments with silver and iodine voltameters yield the following values for the faraday on the two alternative standard conventions regarding atomic weights.

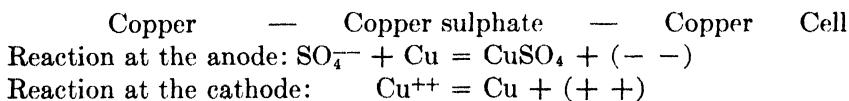
$$F = 96{,}487 \pm 10 \text{ coulombs/gram equivalent (chemical scale)}$$
$$= 96{,}514 \pm 10 \text{ coulombs/gram equivalent (physical scale)}$$

On the chemical scale the mean atomic weight of oxygen is taken as 16, and on the physical scale the atomic weight of the light isotope of oxygen is taken as 16.

An electrolyte may be thought of as a substance whose molecules are held together largely by electrical forces. When an electrolyte is dissolved, the forces binding together the components of the molecule are weakened because of the high dielectric constant of the solvent (κ for water is about 80). This results in a certain fraction of the molecules splitting up into positive and negative ions. Each of these primary ions may attract to itself several solvent molecules which remain bound to it quite firmly and this molecular aggregate drifts slowly through the solution under the influence of the applied field. The negative ions which reach the anode transfer their excess electrons to that electrode and the positive ions are neutralized by the electrons they receive from the cathode. These ions may react with the material of the electrodes or they may simply be liberated or deposited as neutral molecules. For instance, in the case of copper electrodes immersed in a copper sulphate solution

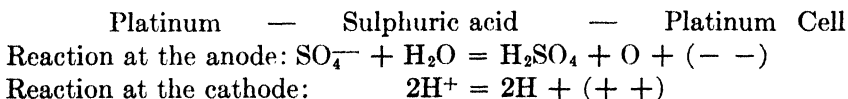
Copper — Copper sulphate — Copper Cell

Reaction at the anode: $SO_4^{--} + Cu = CuSO_4 + (- \ -)$

Reaction at the cathode: $Cu^{++} = Cu + (+ \ +)$

In the case of platinum electrodes in sulphuric acid

Platinum — Sulphuric acid — Platinum Cell

Reaction at the anode: $SO_4^{--} + H_2O = H_2SO_4 + O + (- \ -)$

Reaction at the cathode: $2H^+ = 2H + (+ \ +)$

Not all of the solute molecules are dissociated into their component ions. The process of dissociation is a typical chemical reaction and the degree of ionization is governed by the law of mass action. The resultant conductivity of the solution is determined by the number of ions per unit volume and the rate at which they travel through the solution. Assume, for instance, that a salt is dissolved in water and that there are n' molecules of the salt per unit volume of the solution. Consider for simplicity that the ions are univalent and that dissociation consists in the splitting of the molecule into one positive and one negative

ion, each of charge $\pm e$. Then $n_1' = n_2' = \alpha n'$, where n_1' and n_2' are the numbers of positive and negative ions per unit volume, respectively, and α is the fraction of the molecules dissociated. As Ohm's law is found to hold, the velocity of drift of the ions must be proportional to the electric-field strength. Writing u for the mobility of an ion which is the velocity per unit field, the current density becomes

$$\mathbf{i}_v = (n_1'u_1 + n_2'u_2)e\mathbf{E}$$
$$= \alpha en'(u_1 + u_2)\mathbf{E}$$

The conductivity σ is the ratio $\mathbf{i}_v/\mathbf{E}$, and the equivalent conductivity λ is defined as the conductivity per unit equivalent concentration, *i.e.*,

$$\lambda = \frac{\sigma}{c} = \frac{\alpha n'e(u_1 + u_2)}{m/MV} = \alpha Ne(u_1 + u_2) = \alpha F(u_1 + u_2) \qquad (6.1)$$

The equivalent concentration c of a univalent electrolyte is the number of moles (m/M) per unit volume (V). Thus the equivalent conductivity is proportional to the degree of dissociation and to the sum of the mobilities of the two types of ions. Suitable experiments can be devised to determine the separate mobilities, but these will not be discussed.

FIG. 6.2.—Evolution of oxygen and hydrogen at platinum electrodes by the electrolysis of a sulphuric acid solution.

Varying the concentration affects in general both the degree of dissociation and the mobilities u_1 and u_2. In the case of so-called "weak electrolytes," such as ammonia (NH_4OH) and acetic acid (CH_3COOH), the mobilities are found to depend very little on the concentration and λ may be considered to depend on c only through α. In accordance with the law of mass action the dependence of α on c is given by

$$\frac{\alpha^2}{(1 - \alpha)} = \frac{K}{c}$$

where K is a constant which contains the temperature. When c becomes very small, it is evident that α approaches unity. The limiting conductivity at zero concentration is written λ_0 and it is evident from Eq. (6.1) that for weak electrolytes $\alpha = \lambda/\lambda_0$. By measuring conductivities at decreasing concentrations λ_0 can be found by extrapolation (using the mass-action law), and hence λ can be found at any concentration and the value of the constant K can be determined. In the case of "strong electrolytes" such as KCl, NaCl, $AgNO_3$, etc., dissociation is practically complete at any ordinary concentration (*i.e.*, α is approximately unity) and the variation of λ with concentration is largely due to the

variation in the mobilities of the ions.[1] It may be shown in this case that the equivalent conductivity is proportional to the square root of the concentration, *i.e.*, $\lambda = \lambda_0 - ac^{1/2}$. This theoretical expression, which was discovered empirically by Kohlrausch, has been well verified and may be used to determine λ_0, the equivalent conductivity of the strong electrolyte at zero concentration, by extrapolation. The temperature of the solution also affects the conductivity of the electrolyte. In general it decreases the degree of dissociation and increases the ion mobilities. The net result on the conductivity can be represented roughly for practical purposes by means of a temperature coefficient of 2 per cent. That is, if λ is the conductivity at, say, 18°C., the conductivity λ_T at a temperature T can be written $\lambda_T = \lambda[1 + 0.02(T - 18)]$.

In the preceding discussion Ohm's law has been assumed for electrolytes. That it does hold for similar electrodes in a common solution and in the absence of complicating effects was shown by Kohlrausch. The chief complicating factor is *polarization*. This comes from some inhomogeneity in the electrolyte caused by the ionic motion. As an example consider platinum electrodes in a sulphuric acid solution. When the SO_4^{--} reaches the anode, it gives up its charge and combines with a water molecule to liberate oxygen. At the cathode hydrogen is liberated. The electrodes are surrounded by bubbles of these gases and behave more or less like oxygen and hydrogen electrodes. The potential drop in the surface layer depends on the fraction of the surface effectively covered by these gases which is in turn determined by the nature of the electrode surfaces and the current density. The effect of polarization on resistance measurements may be greatly reduced by the use of alternating current. The familiar Wheatstone bridge is generally used for such measurements; the device for determining the balance condition is usually a pair of headphones. At high frequencies one arm of the bridge must be shunted by an appropriate capacity to neutralize that of the conductivity cell in order to obtain a satisfactory balance (*cf.* Sec. 13.5). The reason for the use of alternating current may be seen from the following approximate analysis: Assuming an ohmic resistance for the solution and that the gas layers give rise to a polarization potential proportional to the total charge q passed, but in the opposite direction, the potential difference between the electrodes is given by

$$V = Ri + Pq$$

where P is a constant of proportionality.

Writing dq/dt for i and assuming an alternating potential $V_0 \sin \omega t$

[1] For the development of the theory of strong electrolytes see Debye and Hückel, *Phys. Zeit.*, **24**, 305 (1923); Onsager, *Phys. Zeit.*, **27**, 388 (1926).

$$R\frac{dq}{dt} + Pq = V_0 \sin \omega t$$

This is of the form of Eq. (C.5) of Appendix C. Substituting the constants in the above equation for those in Eq. (C.6) and neglecting the transient term containing the arbitrary constant α, q becomes

$$q = \frac{V_0}{1 + \frac{P^2}{(R\omega)^2}}\left(\frac{1}{R\omega} \sin \omega t + \frac{P}{(R\omega)^2} \cos \omega t\right)$$

or i, which is equal to dq/dt, is given by

$$i = \frac{V_0}{R} \frac{1}{(1 + (P/R\omega)^2)^{1/2}} \sin (\omega t + \varphi)$$

where $\varphi = \tan^{-1} (P/R\omega)$. If the frequency is made very large, the second term in the denominator becomes negligible. Thus, at high frequencies, Ohm's law ($i = V/R$) is approached. Also, φ approaches zero which means the current is in phase with the potential.

The specific conductance (conductivity) may be deduced from the resistance measurement if the geometrical constant of the cell is known. The following table gives the conductivities of certain representative solutions in reciprocal ohm centimeters or mho/cm. at 18°C for various concentrations.

TABLE I
(Kohlrausch)[1]

Solute	10%	20%	30%	40%	50%	60%	70%	80% (by total wt.)
NaCl....	0.121	0.196						
$ZnSO_4$...	0.032	0.047	0.044					
$AgNO_3$..	0.048	0.087	0.124	0.157	0.186	0.210		
NaOH...	0.309	0.328	0.207	0.121	0.082			
HCl.....	0.630	0.762	0.662	0.515				
HNO_3...	0.461	0.711	0.785	0.733	0.631	0.513	0.396	0.267
H_2SO_4...	0.392	0.653	0.740	0.680	0.541	0.373	0.216	0.111

[1] Multiply entries by 100 to express them in mhos per meter.

6.2. Voltaic Cells.—The name voltaic cell is given to an ordinary electrolytic cell made up of two electrodes dipping in a common electrolyte if a potential difference is thus spontaneously created between the electrodes. The electrodes need not be immersed in the same electrolyte, but they may dip into different solutions separated by some partition permeable to the ions. This difference of potential is produced by forces which are essentially of a nonelectrical nature. The electrical

energy expended in an external circuit connected to the cell terminals has its origin in these forces. They are atomic or electronic forces and may be only partially understood by analogy with ordinary mechanical forces, for they appear to be fundamentally different in character. An account of these forces in atomic terms cannot be undertaken here, but they are the cause of many of the phenomena discussed in the present chapter. It is through them as well as the coulomb force, and the electromagnetic forces that will be introduced later, that energy is transferred between its various forms: electrical, mechanical, chemical, thermal, and radiation.

Voltaic cells may be thought of as devices for transforming chemical into electrical energy. The source of energy is the chemical "heat of reaction"; it gives rise to the potential difference between the cell terminals. Of course, if a current is drawn from the cell this potential difference is altered. It is first of all reduced by a quantity iR_c, where R_c is the resistance of the cell. It is also influenced by any polarization effects that may occur. Cells which are designed for the production of current are of such a nature that polarization does not occur or they contain a suitable depolarizing substance which reacts with the gases liberated at the electrodes. The true potential developed by a cell can be measured most simply if no current is drawn during the process. Thus a potentiometer is particularly suitable for such measurements.

FIG. 6.3.—Sulphuric acid concentration cell.

Concentration Cell.—The diffusion of one gas or vapor into another is a familiar physical phenomenon. If the concentration in an electrolyte is greater in one region than in another, the ions will diffuse from the concentrated to the dilute regions till equilibrium is established. This diffusion process may be used to produce a potential difference between two electrodes in regions of different ionic density. Such a device is known as a concentration cell. Its action depends on a different rate of diffusion for the two types of ions; if they diffuse at the same rate, no potential difference is established between regions of different density. Let $-D$ be the rate of diffusion of ions per unit area per unit concentration gradient (a negative sign is used because the ions move in the direction of decreasing concentration), n the number of ions per unit volume, u the mobility, e the charge per ion, and $\mathbf{E}$ the electric field. Using the subscripts 1 and 2 to distinguish positive and negative ions, then at the start when the densities of positive and negative ions at a given point are equal, the current may be written

$$\mathbf{i}_v = -e(D_1 - D_2)\,\mathbf{grad}\,n + e\mathbf{E}n(u_1 + u_2)$$

Or, since $(\mathbf{grad}\,n)/n = \mathbf{grad}\,(\log_e n)$,

$$\mathbf{i}_v = en(u_1 + u_2)\left[\mathbf{E} - \frac{(D_1 - D_2)}{(u_1 + u_2)}\,\mathbf{grad}\,(\log_e n)\right]$$

The second term in the bracket is the nonelectrical force per unit charge which is brought into existence by the concentration gradient. The first term is the conductivity [compare Eq. (6.1)]. The above equation is an example of the general ohmic law

$$\mathbf{i}_v = \sigma(\mathbf{E} + \mathbf{F}_c) \tag{6.2}$$

where $\mathbf{F}_c$ is a nonelectrical force which, however, produces the motion of charges from one point to another, *i.e.*, an electric current. In electrical equilibrium $\mathbf{E} = -\mathbf{F}_c$ and there is no current flow. If electrodes are placed at positions a and b and the concentrations at these points are n_a and n_b, the potential difference between the electrodes is given by

$$V = V_b - V_a = -\int_a^b \mathbf{F}_c\,\mathbf{dl} = \frac{D_1 - D_2}{u_1 + u_2}\log_e \frac{n_a}{n_b}$$

If the D's are the same or the n's are the same, no potential difference exists. A general statistical mechanical relation exists between D and u for an atomic particle (Sec. 8.1)

$$D = \frac{kT}{e}u \tag{6.3}$$

where T is the absolute temperature and k is a constant (Boltzmann) with the value 1.37×10^{-16} erg or 1.37×10^{-23} joule per degree centigrade per atom. Thus the potential difference produced by the difference in concentration can be written

$$V = \frac{kT}{e}\frac{u_1 - u_2}{u_1 + u_2}\log_e\left(\frac{n_a}{n_b}\right)$$

To obtain an idea of the order of magnitude of this potential difference, $kT/e = 0.026$ volt for 23° C. The ionic concentration gradient cannot be maintained indefinitely and eventually the potential of the cell will drop to zero.

Chemical Cell.—In general, when an electrode is immersed in a solution, a potential difference develops between the two. In order to detect this potential difference, a second electrode must be inserted, but if it is of the same material and the electrolytic solution is homogeneous, the conditions at each electrode are the same and there is no net potential difference between them. On the other hand, if the materials of the electrodes are dissimilar, a potential difference is in

general developed. These phenomena were accounted for by Nernst on the basis of a tendency for the atoms of the electrode to go into solution in the form of ions. This tendency is called a *solution pressure*. If it is positive, the atoms of the electrode tend to go into solution as positive ions; if it is negative, positive ions tend to deposit on the electrode and lose their charge. In the first case, the electrode acquires a negative charge and in the second, a positive one. If the electrodes are insulated the electrostatic forces thus brought into play counteract the solution pressure and limit the amount of metal that goes into solution. Solution pressure is very similar to the more familiar osmotic pressure. The face of the electrode may be thought of as a semipermeable membrane for ions; the metal and the solution form the media on either side.

Assuming that to a first approximation the pressure of the ions is given by the perfect-gas law ($pV = RT$) the work done on allowing a mole of the substance to deposit is given by

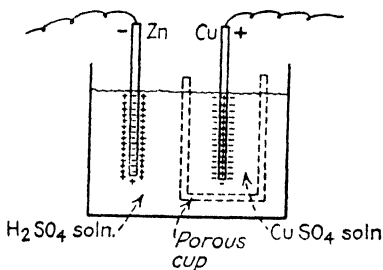



FIG. 6.4.—Chemical cell, Daniell type.

$$W = \int_p^P p\,dV = -\int_p^P \frac{RT\,dp}{p} = -RT \log_e \left(\frac{P}{p}\right)$$

where p and P are the osmotic pressure of the ions in solution and the solution pressure, respectively. This must equal the electrical work done, which is the product of the charge per mole and the potential difference between the solution and electrode; hence

$$V' = -\frac{RT}{Fv} \log_e \left(\frac{P}{p}\right)$$

where v is the charge per ion or valence and V' is the *electrode potential*. Putting in numerical values (R = 8.314 joules/°C. mole) for 18°C.

$$V' = -\frac{0.0251}{v} \log_e \frac{P}{p}$$

The variation in electrode potential with concentration is due to the variation in p. In dilute solutions, to which the perfect-gas law applies, p is proportional to the concentration. Actually, of course, the measurable quantity is the potential difference between two electrodes. Hence, it is a great convenience to choose one standard electrode and refer all other electrode potentials to it. For this purpose the so-called "hydrogen electrode" is chosen. It consists of a strip of platinum covered with a thin layer of platinum black (which occludes hydrogen) saturated with hydrogen at atmospheric pressure. The upper portion of the strip is

surrounded by an atmosphere of hydrogen and the lower portion dips into a solution of unit hydrogen ion "activity." The activity may be thought of as the effective pressure, *i.e.*, the pressure as calculated from the concentration by means of the gas law corrected for any deviation from the latter. Table II gives the sign and magnitude of the potential of an electrode in equilibrium with an aqueous solution of its ions at unit molal activity and at a temperature of 25°C. with respect to the standard hydrogen electrode.

TABLE II

Electrode Reaction	Potential, Volts
$Li - e^* = Li^+$	−2.96
$K - e = K^+$	−2.92
$Na - e = Na^+$	−2.72
$Mg - 2e = Mg^{++}$	−1.55
$Zn - 2e = Zn^{++}$	−0.76
$Fe - 2e = Fe^{++}$	−0.44
$Cd - 2e = Cd^{++}$	−0.40
$Co - 2e = Co^{++}$	−0.29
$Ni - 2e = Ni^{++}$	−0.23
$Sn - 2e = Sn^{++}$	−0.14
$Pb - 2e = Pb^{++}$	−0.12
$Fe - 3e = Fe^{+++}$	−0.04
$H_2 - 2e = 2H^+$	0.00
$Cu^{++} + 2e = Cu$	0.34
$I_2 + 2e = 2I^-$	0.54
$Ag^+ + e = Ag$	0.80
$\frac{1}{2}Hg_2^{++} + e = Hg$	0.80
$Br_2 + 2e = 2Br^-$	1.07
$Cl_2 + 2e = 2Cl^-$	1.36
$Au^{+++} + 3e = Au$	1.36
$F_2 + 2e = 2F^-$	1.90

* e is used as the symbol for an electron.

This table has many important applications, of which only a few will be mentioned. In the first place, the algebraic difference between the listed potential for two electrodes gives approximately the emf. that is developed by a cell corresponding to the specifications. For instance, if a zinc electrode is immersed in a zinc sulphate solution of unit activity separated by a porous partition from a unit active solution of copper sulphate in which is immersed a copper electrode, the potential difference developed between the copper and zinc is

$$0.34 - (-0.76) = 1.10$$

volts. The Daniell cell, which is widely used, corresponds roughly to these specifications. If the two electrodes are joined by an external circuit, a current flows in it from the copper to the zinc. This may be restated by saying that the electrons liberated in the zinc-electrode

reaction flow through the circuit and neutralize the positive ions of copper at that electrode. Eventually all the chemical energy of the cell is delivered as electrical energy to the external circuit; the zinc electrode completely wastes away or the emf. of the cell drops to zero owing to the establishment of chemical equilibrium. In this type of cell, which is shown in Fig. 6.4, polarization is relatively unimportant, for a gas can be liberated only if the electrodes are impure, *i.e.*, if there is no zinc for the SO_4^{--} to combine with in some small locality on the surface of the electrode. Such impurities also give rise to the phenomenon of "local action" which is essentially a small electrochemical circuit on the surface of an electrode.

It may be deduced from Table II that if an electrode is inserted in a solution containing ions of a substance standing below it, the atoms of the electrode will go into solution and the other ions will deposit out. Thus if a zinc electrode is placed in a copper sulphate solution, zinc ions will replace those of copper until the electrode acquires a copper coating. From the position of gold in the table it is evident that it will deposit on any of the other metals listed. The position of an entry with respect to hydrogen determines whether it will dissolve in an acid solution with the simple replacement of hydrogen on the electrode surface. The table also indicates the possibility of electrolytic separation of elements from a common solution. Consider a solution containing copper and silver ions in which two electrodes are inserted. As the emf. applied to them is increased, a point will be reached at which the silver ions are deposited on the anode and this will occur at a potential $0.80 - 0.34 = 0.46$ volt below that at which the copper will leave the solution. The value of the potential at which electrolysis starts depends, of course, on the nature of the negative ion present, and the range of potential in which separation will take place is affected by the concentrations of the metallic ions. However, a practically complete separation of these particular metals can be made in this way. The minimum voltage at which electrolysis between two electrodes takes place is known as the decomposition voltage. It depends on the nature of the electrode surfaces and the many factors that can effect them differentially. It can be determined from Table II in the simpler cases, but when, for instance, a gas is evolved at either electrode the situation is much more difficult to analyze and the previous simple discussion does not apply.

Thermodynamic Theory.—The general thermodynamic theory of a primary cell is of interest. Chemical cells may be divided into two categories: reversible and irreversible. A reversible cell is one in which, if an electric current is sent through it in the sense opposed to the emf. of the cell, the chemical action is reversed. In an irreversible cell the chemical action produces a precipitate or evolves a gas and the process cannot be reversed by sending a current through it in the opposite direction. Strictly speaking, no process is reversible in the sense of reconversion of energy with no loss. In the case of an electric circuit there is always the joule heating (i^2R), but this can be made negligible in comparison with processes depending on the first power of i by keeping the current very small. The Gibbs-Helmholtz equation, which is a

consequence of the fundamental laws of thermodynamics can be applied to a reversible cell. It may be written[1]

$$E = \psi - T\left(\frac{\partial \psi}{\partial T}\right) \tag{6.4}$$

where E is the internal energy, ψ is the "free energy," and T is the absolute temperature. The change in internal energy of the cell for the transfer of a mole of ions is Q, the chemical heat of reaction in joules per mole. The change in available or free energy is given by the product of the total charge transferred and the potential of the cell or $Nev\ V$, where e is the charge of an electron, v the valence of the ion and N the number of ions per mole. But $Ne = F$, the faraday and v and F are constant so Eq. (6.4) becomes

$$Q = vF\left(V - T\frac{dV}{dT}\right)$$

for a chemical cell. This is a very important equation relating the chemical quantity Q and the cell potential and temperature. It is chiefly of use in calculating Q from measurements of the other quantities that appear. A further discussion of this equation together with examples of its verification and use will be found in treatises on physical chemistry.

Practical Voltaic Cells.—Though the commercial production of electric current is carried out entirely by means of electromagnetic generators chemical cells still have important uses. The Daniell cell which is depicted in Fig. 6.4 has an amalgamated zinc cathode immersed in a 20 per cent H_2SO_4 solution and a copper anode in a saturated solution of $CuSO_4$. The emf. developed is 1.06 volts. A more convenient form of cell for most purposes is the "dry cell." A cylindrical zinc case which forms the cathode is packed with moist NH_4Cl. The anode is a carbon rod at the center surrounded with MnO_2 which reacts with the hydrogen liberated at that electrode and is the depolarizing agent. Sufficient moisture is present for electrolytic conduction. The top of the cell is sealed with a resinous compound, and as a whole the cell is portable, rugged, and dependable. The emf. developed is 1.53 volts when new.

Special types of cells, which are carefully designed and constructed of suitable materials, will supply an exceedingly constant emf. if the currents drawn are limited to very minute values. Their reproducibility and the constancy of the emf. developed is such that certain of these cells have been adopted as secondary standards of potential. The type in most general use at present is the Weston standard cell which is shown diagrammatically in Fig. 6.5. The cathode is a cadmium amalgam covered with a layer of $CdSO_4$ crystals. The electrolyte is a saturated solution

[1] See any thermodynamics text.

of $CdSO_4$. The anode is mercury which is covered with a Hg_2SO_4-$CdSO_4$ paste which is in turn covered by a layer of $CdSO_4$ crystals. The emf. developed at 20°C. is 1.0187 volts; its variation with temperature is given by the following expression:

$$V_T = 1.0187[1 - 0.000037(T - 20)]$$

The cell is hermetically sealed to eliminate many sources of variation. When new it is accompanied by a certificate of standardization; aging over a few years will produce changes in the emf. of only a few parts in 100,000 if it is carefully used in suitably protected potentiometer circuits. The internal resistance is low and the effects of polarization are negligible for the minute currents that are drawn. For the most accurate work a number of cells should be kept in use and compared with one another from time to time. Also comparison with the primary standards of resistance and current must be made at regular intervals.

FIG. 6.5.—Weston normal standard cell.

In addition to primary cells there are certain so-called secondary or storage cells which are of commercial importance. These are reversible to such an extent that in practice the cells are charged by sending a current through them in such a sense as to reverse the normal chemical reaction responsible for the emf. of the cell. The commonest is the lead cell which has a Pb cathode and a PbO_2 anode immersed in an H_2SO_4 solution with a specific gravity of 1.25. In order to decrease the internal resistance so that large currents may be drawn a number of plate electrodes are arranged in parallel separated by wooden spacers but in the same solution. Such a battery is shown in Fig. 6.6. The cathodes are lead plates and the anodes are generally made by casting a PbO_2 paste in skeleton lead forms. During the discharge process SO_4^{--} reacts at the Pb plates to form $PbSO_4$, and at the anode the H^+ probably reduces the PbO_2 to PbO with the formation of water, and the PbO reacts with H_2SO_4 to form more water and further $PbSO_4$. The charging process is the reverse and all the $PbSO_4$ disappears. The acid density falls as water is formed by the discharge process and when it reaches about 1.1, the cell should be recharged. A hydrometer must be used to indicate the state of the cell, as the emf. changes very little from the charged to

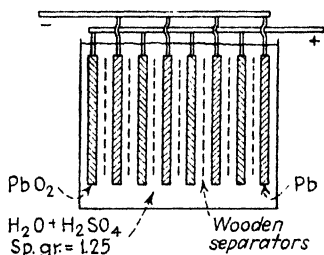



FIG. 6.6.—Lead storage cell.

the discharged condition. When fully charged, the emf. is about 2.05 volts, when completely discharged it is of the order of 1.80 volts. The internal resistance of a battery may be made very low, of the order of 0.01 ohm, and there is no appreciable polarization. The capacities of cells vary from 10 to 1,000 amp.-hr., *i.e.*, from 3.6×10^4 to 3.6×10^6 coulombs. The only attention the cells need is the addition of distilled water from time to time to replenish that lost by evaporation. The chief disadvantage of these cells is their great weight. A lighter cell may be made with iron and nickel oxide electrodes in a 20 per cent KOH solution. This is known as the Edison cell and it has the further advantages of being more rugged and of not deteriorating when standing in a discharged condition. However, it has the great disadvantage that its emf. varies widely with the state of charge. Its maximum value at full charge is about 1.4 volts.

6.3. Electron Structure of Crystals.—In order to present a more unified picture of the electrical properties of matter in the solid state it is necessary to extend somewhat the very elementary description of solids in terms of the atoms composing them that was given in Sec. 1.1. This account will necessarily be restricted to a simple qualitative presentation of those features which are necessary to understand the more salient phenomena with which the present chapter is concerned.[1] Crystals are composed of atoms or molecules associated together in a regular spatial lattice array. The particular configuration assumed by these atoms is presumably determined by the equilibrium of the various forces of electrical and nonelectrical natures concerned in atomic interactions. The electrons that are most firmly bound to the individual atoms are but little influenced by the presence of the neighboring atoms. On the other hand, electrons whose binding energy is not great in comparison with the interaction energy of the atoms in the crystal are profoundly affected by the presence of adjacent atoms. The phenomena with which we are here concerned are interpretable in terms of the behavior of these more loosely bound or valence electrons, and a brief account will be given of their properties on the basis of elementary quantum-theory concepts.

The outstanding feature of atomic structure is the discrete set of energy states or levels that can be occupied by the electrons of a single isolated atom. These states are very precisely determined, and a change in the atom from one state to another is accompanied by the absorption or emission of the energy difference between them as exhibited by atomic line spectra. The description of atoms in terms of these energy levels is merely another way of specifying the motion of the electrons composing

[1] General references: MOTT and JONES, "Properties of Metals and Alloys," Oxford University Press, New York, 1936: SEITZ, "Modern Theory of Solids," McGraw-Hill Book Company, Inc., New York, 1940.

their structure. A basic law of this structure is that no two electrons can be specified in precisely the same way in terms of the series of parameters necessary to characterize their motion. This is known as the *exclusion principle.* When atoms are brought together to form a crystal, the aggregate may be thought of in the same terms as for the individual atoms. The total number of energy levels is the sum of the energy levels for each atom, but their actual magnitudes are altered by the spatial proximity of the atoms and consequent energy of interaction between them. The vast number of levels thus brought into existence by bringing, say, 10^{20} atoms together to form a crystal block renders the density of levels in certain energy intervals so great that for practical purposes it

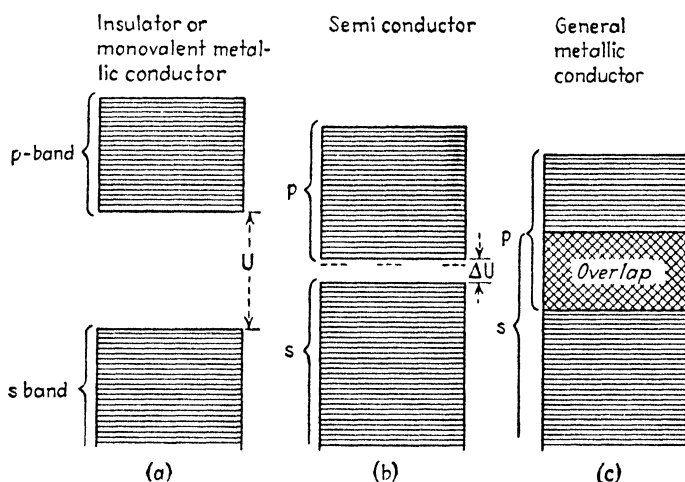



FIG. 6.7.—Schematic representation of three different types of electron band structure.

is essentially a continuum over these intervals. Such groups of level continua are known as *energy bands.* The bands arising from atomic energy levels of different types may or may not overlap, depending on the particular atoms concerned and their spacing in the crystal. The electrical properties that the crystal exhibits can be described most succinctly in terms of this band structure and the number of electrons that occupy it under the limitations on their motion imposed by the exclusion principle.

Figure 6.7 illustrates schematically the three types of situation that may arise in the formation of bands. The designations s and p for the lowest and next lowest bands are carried over from atomic nomenclature in which these letters refer to electronic angular momenta in the group of low-energy states. If the band resulting from the coalescence of s states is separated by an energy of several electron volts from that arising from the p states, there is little likelihood that the distribution of thermal energy among the s electrons will enable any appreciable number of them

to achieve entrance to the p band and the intervening region is excluded from occupancy by electrons. Two cases may then be distinguished. The first is that in which all the levels in the s band are occupied by electrons; and as there are two atomic levels per atom, this is the situation when the atoms are divalent, *i.e.*, have two electrons apiece. If all s levels are occupied, the direction and magnitude of the motion of each electron in the crystal are completely determined by the exclusion principle. The application of an electric field cannot alter this isotropic distribution of electron velocities. Hence there can be no net motion of charge in one direction under the influence of the field, and the crystal is said to be an insulator. On the other hand, if the atoms are monovalent, *i.e.*, have but one free s electron apiece, only half of the s levels are occupied and the application of a field can induce electrons to occupy those levels in the band corresponding to a net motion in the direction determined by the field. Thus a current flows, and the crystal is a conductor. Illustrations would be the monovalent metals lithium, sodium, copper, silver, etc. If the particular atomic energy levels and spacing in the crystal are such that the bands composed of s and p levels overlap [(*c*) of Fig. 6.7] there are available levels corresponding to electron paths having a net preponderance in the direction determined by the field, even in the case of divalent elements and conduction results. Illustrations are the divalent metals such as magnesium, calcium, zinc, and mercury.

In the case that the separation between the s and p bands is quite small, ΔU of Fig. 6.7*b*, the crystals are said to be *intrinsic semiconductors* (Secs. 5.2 and 5.4). When this forbidden energy range is of the order of a few tenths of an electron volt, thermal energies are significant in raising electrons from the full s band to the p band and the conductivity is very temperature dependent. The presence of foreign atoms in a crystal lattice may greatly increase the conductivity of an insulator, and such a material is said to be an *impurity semiconductor*. The dashes between the levels indicate schematically possible local impurity electron levels due to foreign atoms or imperfections in the lattice. These do not represent electrons that can move freely through the crystal, but they play a significant role in conduction phenomena. If these local levels are such that they can receive electrons, for example isolated positive ions imbedded in the lattice, and if they are not very high in energy above the top of the s band, they may gain occasional electrons that by virtue of their thermal energy are able to reach them. This process, by removing some electrons from the full s band, relaxes the conditions on the electron velocities and permits a net flow in one direction or conduction. Such materials are known as *P-type* semiconductors, as the carriers in the s band behave in some ways (Hall effect) as if they were positive charges. If the local levels on the other band are but little below the p band in

energy and are such as to have an excess of electrons associated with them, they can contribute these electrons to the p band on the absorption of thermal energy, and conduction in that band results. Such impurity semiconductors are known as *N-type*, as the carriers behave consistently as if they were negatively charged.

The way in which the conduction electrons distribute themselves in a band continuum is primarily a function of the temperature, although it also depends to some degree on the number of electrons per unit volume and nature of the levels in the band. It can be shown on the basis of the statistics that electrons are known to obey that the number n' of electrons per unit energy interval per unit volume of the crystal is given by

$$n'\,du = C\frac{\sqrt{u}\,du}{e^{\left(\frac{u-u'}{kT}\right)}+1} \tag{6.5}$$

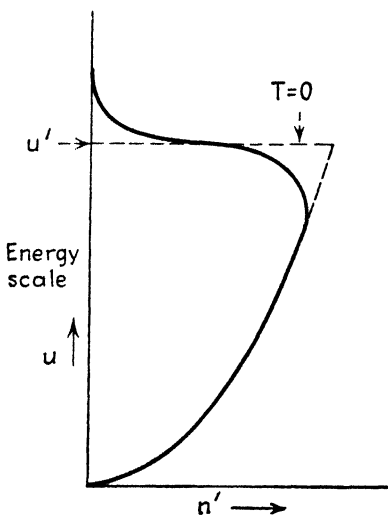



FIG. 6.8.—Plot of Eq. (6.5) for $T = 0$ and a high temperature.

Here u is the energy coordinate, C and the characteristic energy u' are functions of the number of electrons per unit volume and the nature of the band structure, k is Boltzmann's constant, and T is the absolute temperature. A representative graph of Eq. (6.5) is shown in Fig. 6.8 for $T = 0$ and a large value of T. From the change in the form of the curve with T it is clear that higher energy levels tend to become occupied at the expense of slightly lower levels as T increases. There is, however, little change in the occupancy of the lowest energy levels or in the total electronic energy with temperature. The average energy of an electron at moderate temperatures can be calculated from Eq. (6.5) to be[1]

$$\bar{u} = \frac{3}{5}u_0'\left[1 + \frac{5\pi^2}{12}\left(\frac{kT}{u_0'}\right)^2\right] \tag{6.6}$$

where u_0' is written for $\frac{h^2}{2m^*}\left(\frac{3n''}{8\pi}\right)^{2/3}$, n'' is the number of electrons per unit volume, and m^* is the apparent mass of the electron as influenced by the lattice and electron structure forces. In case gaps exist in the band structure, as in Fig. 6.7 (*b*), their influence on the constants in Eq. (6.5) is very pronounced. However, if one is concerned principally with

[1] SEITZ, *loc. cit.*

electrons near the top of the energy distribution, *i.e.*, for which u is close to u', the exponential in the denominator is the dominant factor through which n' is influenced by T. If a gap in the allowed energy levels of width Δu occurs just above u' an increase in temperature enables some electrons to occupy levels immediately above it at the expense of those immediately below. The distribution of these electrons in energy is exponential about a mean energy $u'' = u' + \Delta u/2$. In consequence the number of electrons free for conduction is proportional to $e^{-\Delta u/2}$ where Δu is the smallest energy gap in the general or local band structure that must be traversed by electrons in order that they may move freely through the structure. Thus the conductivity should vary exponentially with the temperature, becoming greater as the energy gap that must be traversed becomes smaller. This is in accordance with the semiconductor phenomena discussed in Sec. 5.4.

The boundary or junction between two different crystalline materials may be thought of as a membrane separating two different concentrations of electrons. The mean energy of the electrons on the two sides is given by Eq. (6.6) in terms of these concentrations, and the nature of the levels in the bands through the dependence of u on n' and m^*. Thus the situation resembles somewhat that of the concentration cell. In equilibrium the concentrations adjust themselves in such a way that no net work would be done in taking an average mobile electron from one side to the other. Any electrical forces that may exist are exactly counterbalanced by the nonelectrical forces arising from the electron structures of the two metals. If an emf. is applied across the junction and electrons move from one side to the other, work may be done by or against the nonelectrical band structure forces, as $\bar{u}$ of Eq. (6.6) will in general not be the same in the two metals. If the electron motion across the boundary is not constrained by available levels on either side, the situation is quite symmetrical and no difference in conductivity in the two senses would be anticipated. The situation is quite different at a junction if the band structure influences the flow of current across the boundary. Figure 6.9 illustrates schematically the conditions at a semiconductor-metal junction. The line u'', known as the *Fermi level*, represents the mean energy of those electrons which are free to move through the lattice at the temperature T. Its position relative to the semiconductor levels is determined by the condition that no net work would be done in transferring an average electron near the top of the distribution in the metal, *i.e.*, an electron capable of motion under the exclusion-principle restrictions, from one side to the other. In the case of an intrinsic semiconductor, which is one having no local impurity levels, the average energy of the conduction electrons is about midway in the forbidden band and as many electrons populate the lowest p levels as have left the uppermost

s levels at the temperature T. In the P-type semiconductor the thermal equilibrium is between the local acceptor levels and the upper s levels, so u'' occurs about halfway between these levels. For analogous reasons it occurs midway between donator levels and the bottom of the p band in the case of a junction between a metal and an N-type semiconductor. These asymmetries account for asymmetrical conduction or rectification at such junctions. If an emf. is imposed at a P-type junction in such a sense as to lower u'', many levels are available for electrons in the semiconductor to occupy if they move over into the metal and a large current may flow. If the emf. is applied in the opposite sense, only those few electrons in the upper tail of the temperature distribution are differentially affected and permitted to flow into the semiconductor. By an analogous argument it is clear that the opposite effect would be anticipated at an N-type junction. This description of the behavior of the electrons in terms of the structures on the two sides of the boundary accounts for the general features of rectification mentioned previously in Sec. 5.2.

FIG. 6.9.—Schematic illustration of conditions at representative semiconductor-metal junctions.

6.4. Thermoelectric Effects. *Seebeck Effect.*—In 1821, Seebeck made the observation that if two wires of different metals are joined at their ends to form a conducting circuit, a current will flow around it if the two junctions are maintained at different temperatures. Such a circuit of two metals A and B with the junctions at the temperatures T_1 and T_2 is shown in Fig. 6.12. The electromotive force in the circuit depends on the two metals employed and the temperatures of the junctions. The circulating current depends, of course, on the resistance as well. The current can be detected with a suitable meter, but as it depends on the resistance which is itself a function of the variable temperature throughout the circuit, the emf. developed is the more significant quantity and it is generally measured with a potentiometer.

Peltier Effect.—In 1834, Peltier observed the inverse effect, namely, that when a current flows across the junction between two different metals heat is either generated or absorbed. This is not to be confused

with the joule heating which is a generation of heat proportional to i^2, for the Peltier effect may be either heating or cooling, depending on the sense of passage of current across the boundary, and it is proportional to the first power of the current. It may be demonstrated with the apparatus indicated diagrammatically in Fig. 6.10. Two bulbs containing air are joined by a capillary tube in which is a drop of mercury. One junction of the circuit is in each bulb. When the current is sent in one direction the junction a is heated and b is cooled, thus the air surrounding a expands and that around b contracts forcing the droplet to the right. When the sense of the current is reversed by means of the switch, the droplet moves in the opposite direction. The Peltier effect also depends on the materials of the wires A and B and on the temperatures.

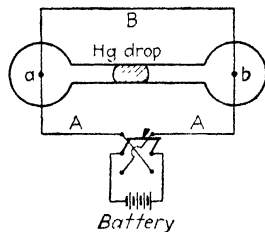



FIG. 6.10.—Apparatus for demonstrating the Peltier effect.

Thomson Effect.—The third associated effect was pointed out by Thomson (Lord Kelvin) in 1851. If a temperature gradient exists in a conductor, an electric potential gradient is also brought into existence. The necessity for this may be shown on thermodynamic grounds. The effect may be demonstrated with the apparatus shown schematically in Fig. 6.11. A fine wire w is soldered to the two limbs of a heavy copper yoke. The center of the wire C and the center of the yoke are joined through a circuit containing a galvanometer and key. When a battery is connected to the limbs of the yoke, the whole forms a bridge circuit which may be balanced by means of the variable resistance S, shunting one arm of the yoke. The fine wire should be well insulated thermally from its surroundings. The bridge is balanced after the wire has reached its equilibrium temperature with the current flowing through it from A to B. The switch is then reversed so that the current flows from B to A and the bridge is found to be out of balance. The reason for this is the temperature gradient and hence the potential gradient existing in the wire. In the two sections of the wire the thermal gradient has the directions $C \rightarrow A$ and $C \rightarrow B$, owing to the cooling effects of the limbs of the yoke. For the first balance condition the ratio V_{AC}/V_{CB} is that of the potential drops in the corresponding arms of the yoke which is a constant and may be written K. But $V_{AC} = iR_{AC} - e$ and $V_{CB} = iR_{CB} + e$, where e is the emf. introduced in the direction of the thermal gradient. For the bridge to remain in balance with the sign of the current reversed the following must evidently be true:

FIG. 6.11.—Apparatus for demonstrating the Thomson effect.

$$\frac{iR_{AC} - e}{iR_{CB} + e} = \frac{-iR_{AC} - e}{-iR_{CB} + e} = K$$

This equation can be true only if $e = 0$; therefore the unbalance of the bridge produced by reversing the current shows that a finite emf. e is brought into existence by the temperature gradient in the wire.

These effects are to be expected in terms of the discussion in the preceding section. Consider the thermoelectric circuit composed of metals A and B, shown in Fig. 6.12. Let the *Peltier coefficient* ${}_T\Pi_{AB}$ be the work done against the nonelectrical forces when a number of electrons equivalent to a unit charge is taken across the junction from metal A to metal B

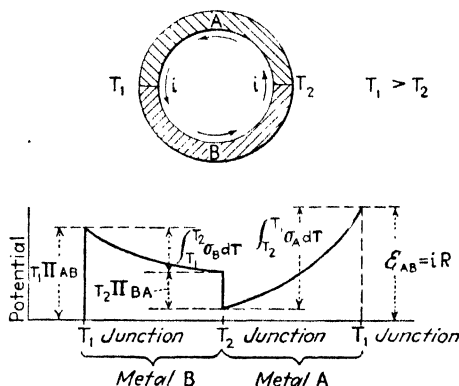



FIG. 6.12.—Analysis of the thermoelectric circuit.

at the temperature T. Let the work done against the nonelectrical concentration gradient forces in moving these electrons through a temperature difference dT in a metal be $\sigma\,dT$, where σ, which depends on the metal, is known as the *Thomson coefficient*. Then if the emf. existing in the circuit of Fig. 6.12 is $\mathcal{E}$, the equation stating that the work done against the forces on the electrons in the materials in circulating a unit charge must be derived from the electrical forces is

$$_{T_1T_2}\mathcal{E}_{AB} = {}_{T_1}\Pi_{AB} + \int_{T_1}^{T_2} \sigma_B\,dT - {}_{T_2}\Pi_{AB} - \int_{T_1}^{T_2} \sigma_A\,dT \tag{6.7}$$

The derivation of this equation is indicated graphically in Fig. 6.11. This is the fundamental equation of the thermoelectric circuit.

The circuits used in practice are seldom as simple as that of Fig. 6.12. In order to measure the emf., for instance, the circuit must be broken and a meter introduced. A consideration of the three element circuit of Fig. 6.13 will suffice to establish the general laws of a more complex circuit. Analyzing this circuit in the same way as that of Fig. 6.12 the emf. developed in it is seen to be

$$_{T_1T_2T_3}\mathcal{E}_{ABC} = {}_{T_2}\Pi_{BA} + \int_{T_2}^{T_1} \sigma_A\,dT + {}_{T_1}\Pi_{AC} + \int_{T_1}^{T_3} \sigma_C\,dT + {}_{T_3}\Pi_{CB} + \int_{T_3}^{T_2} \sigma_B\,dT$$

Now let C represent the measuring instrument the terminals of which are at the same temperature, say, T_1. Letting $T_1 = T_3$ in the above expression

$$_{T_1T_2T_1}\mathcal{E}_{ABC} = {}_{T_2}\Pi_{BA} + {}_{T_1}(\Pi_{AC} + \Pi_{CB}) + \int_{T_2}^{T_1} (\sigma_A - \sigma_B)\, dT$$

And since at any constant temperature, in this instance T_1

$$\Pi_{AC} + \Pi_{CB} + \Pi_{BA} = 0$$

The expression for the emf. becomes

$$_{T_1T_2T_1}\mathcal{E}_{ABC} = {}_{T_1}\Pi_{AB} - {}_{T_2}\Pi_{AB} + \int_{T_2}^{T_1} (\sigma_A - \sigma_B)\, dT = {}_{T_1T_2}\mathcal{E}_{AB}$$

Thus the emf. is the same as that for the simple circuit composed of the metals A and B with the junction temperatures T_1 and T_2. Therefore the insertion of a conductor with terminals at the same temperature in a thermoelectric circuit has no effect on the emf. developed by the circuit. The terminals of an instrument can easily be kept at the same temperature, but it is seldom convenient to have this one of the reference temperatures. Therefore the standard thermocouple circuit, for example, is of the type of Fig. 6.13 with C of the same material as B and the instrument terminals inserted in the circuit at the junction T_3. The emf. is independent of the instrument temperature T_3.

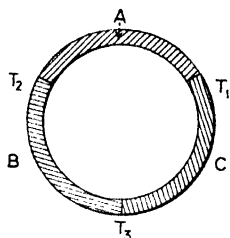



FIG. 6.13.—Three elements in a thermoelectric circuit.

The complete analysis of the thermoelectric circuit to determine the constants introduced in terms of the electromotive force and the temperature requires both Eqs. (6.4) and (6.7). Considering T_1 as a variable temperature T, and T_2 as a constant, Eq. (6.7) may be differentiated with respect to T to give

$$\frac{d\mathcal{E}}{dT} = \frac{d\Pi}{dT} + \Delta\sigma \tag{6.8}$$

where $\Delta\sigma$ is written for $\sigma_A - \sigma_B$ and the subscripts indicative of the metals have been dropped. Consider Eq. (6.4) as applied to the junction at a temperature T. The difference in internal energy for the transfer of a unit charge across the boundary is the difference in the Thomson coefficients (formally similar to specific heats) multiplied by the temperature, *i.e.*, $E = \Delta\sigma T$. The free energy made available by the passage of a unit charge is equal to the potential difference between the two sides of the boundary or Π. Therefore Eq. (6.4) becomes

$$\Delta\sigma\, T = \Pi - T\frac{d\Pi}{dT} = -T^2\frac{d(\Pi/T)}{dT}$$

Eliminating $\Delta\sigma$ between these equations

$$\Pi = T\frac{d\mathcal{E}}{dT} \tag{6.9}$$

and

$$\Delta\sigma = -T\frac{d^2\mathcal{E}}{dT^2} \tag{6.10}$$

Thus the Peltier coefficient and the difference between the sigmas can be calculated from the absolute temperature and the first and second derivatives of the emf. with respect to the temperature. The sigmas themselves cannot be separately determined from these measurements.

It is found experimentally that the emf. in a thermoelectric circuit may be represented very satisfactorily by a quadratic function of the difference between the temperatures of the junctions. As there is, of course, no constant term the empirical equation can be written

$$\mathcal{E} = \alpha t + \tfrac{1}{2}\beta t^2 \tag{6.11}$$

where $t = T - T_0$. α and β are constants characteristic of the metals. The thermoelectric coefficients can be written in terms of α and β. From Eqs. (6.9) and (6.10)

$$\Pi = (\alpha - \beta T_0)T + \beta T^2$$
$$\Delta\sigma = -\beta T$$

The quantity $d\mathcal{E}/dT$ is known as the *thermoelectric power;* it is seen to be equal to $(\alpha - \beta T_0) + \beta T$. It is evident from Eq. (6.11) that $\mathcal{E}$ has an extreme value when $t = -\frac{\alpha}{\beta}\cdot$ It is zero for $t = 0$ $(T = T_0)$ and for $t = -\frac{2\alpha}{\beta}\left(T = T_0 - \frac{2\alpha}{\beta}\right)\cdot$ The extreme value of t is known as the *neutral temperature*, and for the potential to be a single-valued function of the temperature t must always lie on one side or the other of this value. If Eq. (6.7) is written for three pairs of metals AB, BC, CA for any two temperatures T_1 and T_2 and these three equations are added together, the right-hand side is seen to reduce to zero. Therefore

$$\mathcal{E}_{AB} + \mathcal{E}_{BC} + \mathcal{E}_{CA} = 0 \tag{6.12}$$

If two of these quantities are known the third can be determined from this equation. As a consequence it is only necessary to list the constants α and β for the elements in terms of some one standard element. Lead is generally chosen as this standard since its value of sigma is so small as to be negligible. Then sigma for any metal is given by βT, where the values of β are given in the following table. From Eq. (6.12) the con-

stants α_{AB} and β_{AB} for any two metal pairs are given by

$$\alpha_{AB} = \alpha_A - \alpha_B \qquad \beta_{AB} = \beta_A - \beta_B$$

where α_A and β_A are the values against lead listed in the table. A positive sign indicates that the current at the hot junction flows from lead to the metal listed.

Thermocouples are used over a very wide range for temperature measurement. One junction is placed in contact with the body whose temperature is to be measured and the other is maintained at a constant temperature. This reference temperature is generally that of melting ice so that t is obtained directly in degrees centigrade. For accurate work the emf. developed must be measured by a potentiometer and care must be taken to ensure that any additional junction pairs in the circuit are kept at the same temperature. For the low-temperature range (−200 to

TABLE III[1]

Substance	α in 10^{-6} volt/°C.	β in 10^{-8} volt/°C.[2]	Temperature range, °C.
Bismuth (commercial)	−43.688	−46.47	−200 to 100
Constantan (60% Cu − 40% Ni)	−38.105	− 8.88	0 to 400
Copper (hard drawn)	2.76	1.22	0 to 100
Gold	2.90	0.68	−200 to 125
Iron (soft)	16.65	− 2.966	−230 to 100
Manganin (84% Cu, 12% Mn, 4% Ni)	1.366	0.083	0 to 100
Mercury	− 8.8103	− 3.333	0 to 200
Molybdenum	5.892	4.334	0 to 100
Nickel	−19.067	− 3.022	0 to 200
Platinum (Baker)	− 1.788	− 3.460	0 to 100
Platinum-iridium			
85% Pt − 15% Ir	14.083	1.06	0 to 1200
90% Pt − 10% Ir	13.208	0.75	
Platinum-rhodium			
85% Pt − 15% Rh	6.69	1.07	0 to 1600
90% Pt − 10% Rh	7.013	0.64	(against Pt)
Silver (annealed)	2.50	1.15	0 to 100
Tungsten	1.594	3.41	0 to 100

[1] Handbook of Chemistry and Physics.

400°C.) a copper-constantan thermocouple is generally used. For the high ranges chromel-alumel is useful from 0 to 1400°C. and platinum-platinum-rhodium from 0 to 1750°C. A pair of junctions to be used for accurate temperature measurement should be standardized at three temperatures well spaced throughout the range to be used. The two temperature pairs will determine the constants α and β. Useful standardizing temperatures are the following:

Substance	Melting point, °C.	Substance	Boiling point, °C.
Ice	0	Water	100
Lead	327.4	Mercury	356.9
Antimony	630.5	Sulphur	444.6
Silver	960.5	Selenium	688
Copper	1083.0	Zinc	907
Nickel	1455	Lead	1620
Platinum	1773.5		

To measure temperatures above about 1700°C. an optical pyrometer must be used, for most substances with suitable electrical properties melt.

An important use of thermocouples is in the measurement of high-frequency alternating currents. The current is sent through a fine wire in an evacuated chamber and a junction is in thermal contact with the wire somewhere near its center. The joule heating of the wire when a current flows through it raises the temperature of the junction above that of the other pair of terminals which are connected to a direct-current meter and remain practically at room temperature. The meter is calibrated by sending a constant current through the wire. The deflection is approximately proportional to the square of the current heating the wire. The arrangement is indicated diagrammatically in Fig. 6.14*a*.

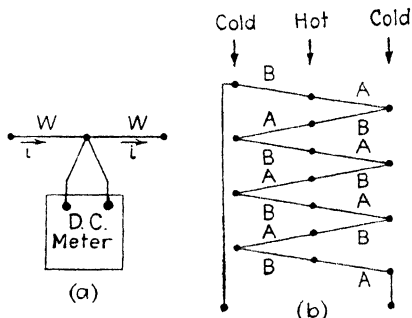



FIG. 6.14.—Circuits visiting thermoelectric junctions. (*a*) Thermocouple meter. (*b*) Thermopile.

A number of junctions arranged in series constitutes a thermopile. The emf. developed is the number of junctions times the emf. per junction, and the resistance is also proportional to the number of junction pairs. Thus it is a more suitable device for working into a high-resistance instrument. It is particularly useful for radiation measurements. The radiation falls on the central junctions of Fig. 6.14*b*, and if they are suitably blackened, a large fraction of this energy is absorbed and the temperature of these junctions rises. The outer junctions are in thermal contact with a relatively large mass of metal and their temperature remains constant. The emf. developed is approximately proportional to the radiant energy falling on the thermopile. In a well-designed thermocouple it is possible to generate 1 μv. per microwatt of radiation.

6.5. Thermionic Emission.—Except to the extent that the Peltier coefficient describes an effect at the boundary of a metal, the preceding

discussion has been limited to phenomena within the body of a metal. However, as electrons populate higher levels in the band structure with increasing temperature, it is evident that at some temperature an appreciable number of these electrons should be able actually to surmount the energy barrier at the surface that retains them within the metal. On starting with a modification of Eq. (6.5) the number of electrons leaving a unit area of the metal surface at the temperature T can be calculated.

In terms of the components of momentum of electrons within the metal the number per unit volume having momenta between p_x and $p_x + dp_x$, p_y and $p_y + dp_y$, and p_z and $p_z + dp_z$ is

$$n'(p)\, dp_x\, dp_y\, dp_z = \frac{2}{h^3} \frac{dp_x\, dp_y\, dp_z}{e^{\frac{u-u'}{kT}} + 1}$$

where h is an atomic constant (Planck's constant) equal to 6.62×10^{-34} joule sec. and $u = (1/2m)(p_x^2 + p_y^2 + p_z^2)$. The number of electrons striking unit area of the crystal surface normal to the x axis per second is this quantity times $v_x = \partial u/\partial p_x$. The number that is able to surmount the surface barrier and leave the crystal is the number for which the energy exceeds that necessary to surmount the barrier by the amount $1/2m(p_y^2 + p_z^2)$, which is unchanged in passing through the barrier. The height of the barrier potential is known as the *work function*, φ, and in terms of it the energy associated with the x component of motion must be equal or greater than

$$u \geqslant e\varphi + \frac{1}{2m}(p_y^2 + p_z^2) + u' \equiv u^*$$

Thus the number of electrons leaving unit area per second times the electronic charge is the thermionic current I per unit area

$$I = \frac{2e}{h^3} \int_0^\infty \int_0^\infty \int_{u^*}^\infty \frac{du\, dp_y\, dp_z}{e^{\frac{u-u'}{kT}} + 1}$$

$$= \frac{2ekT}{h^3} \int_0^\infty \int_0^\infty \log_e \left[1 + e^{\frac{-1}{kT}\left(e\varphi + \frac{1}{2m}(p_y^2 + p_z^2)\right)}\right] dp_y\, dp_z$$

As $e\varphi/kT$ is generally much larger than unity, the exponential term is small and the logarithm can be expanded as simply the exponential term. Also writing $p_x^2 + p_y^2 = r^2$ the double integral can be considered as a single integral

$$\int\int f(r^2)\, dp_x\, dp_y = 2\pi \int f(r^2) r\, dr$$

$$I = \frac{4\pi ekT}{h^3} e^{-\frac{e\varphi}{kT}} \int_0^\infty e^{-\frac{r^2}{2mkT}} r\, dr$$

$$= \frac{4\pi em(kT)^2}{h^3} e^{-\frac{e\varphi}{kT}} \tag{6.13}$$

This is known as the *Richardson-Dushman equation* and was first derived from the application of thermodynamics to the electrons considered as a gas occupying the crystal volume.

The height of the potential barrier φ may be considered to be due to the image force on an electron leaving the plane surface modified possibly by any applied electric field tending to extract the electrons. From

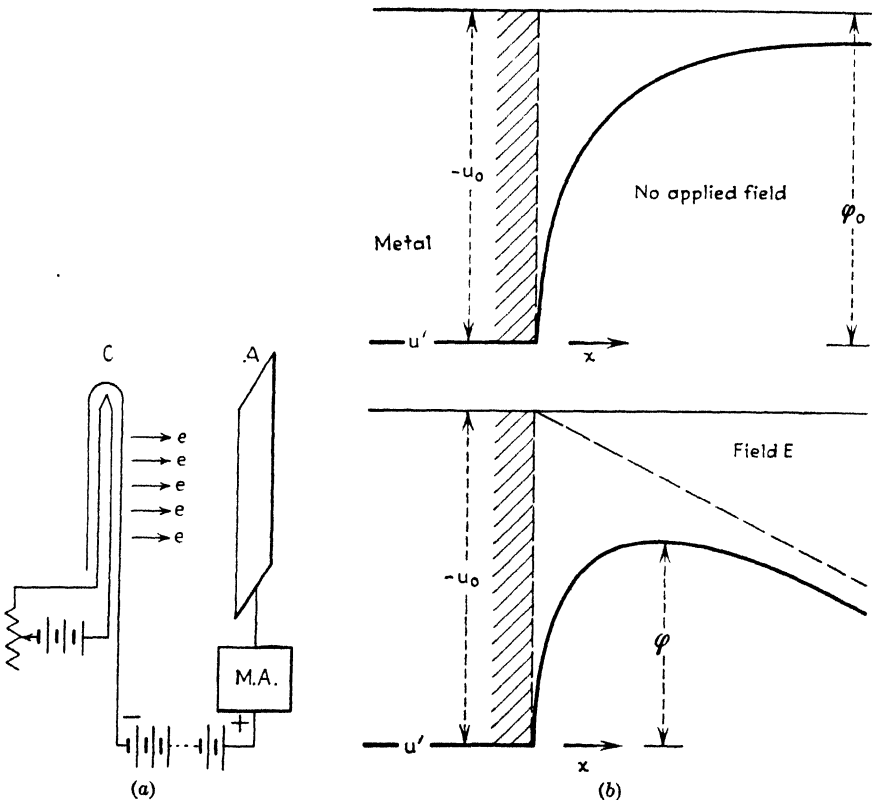



FIG. 6.15.—(*a*) Schematic arrangement for measuring the thermionic emission characteristic. (*b*) Potential function of an electron outside a plane metal surface due to the image force.

Sec. 1.7 the force toward the surface exerted on charge q at a distance x from a plane surface is $q^2/16\pi\kappa_0 x^2$, or the potential associated with this force on an electron is $(\varphi_0 e - e^2/16\pi\kappa_0 x)$, to agree with the notation of Fig. 6.15*b* when x becomes very great. If an electric field is also applied, its potential may be represented by $-Eex$. Thus the total potential is $e\varphi_0 - e\left(Ex + \frac{e}{16\pi\kappa_0 x}\right)$, and this has a maximum at $x = \left(\frac{e}{16\pi\kappa_0 E}\right)^{1/2}$ at which its value is the effective barrier height φ by definition or

$$\varphi = \varphi_0 - \left(\frac{e}{4\pi\kappa_0}\right)^{\frac{1}{2}} E^{\frac{1}{2}} \tag{6.14}$$

The variation of φ and consequent variation of I with E are known as the *Schottky effect.* Although the factor by which $E^{\frac{1}{2}}$ is multiplied is very small, it is not necessary that the second term actually reach more than a small fraction of the magnitude of φ_0 to influence the emission markedly. The Schottky effect or field emission, as it is sometimes called, can produce measurable currents at gradients of 10^6 volts per centimeter even from smooth clean surfaces. The only correction to Eq. (6.13) to the accuracy of the present analysis is to remark that no allowance has been made for the possibility that the energy condition may not be the only one retaining an electron in the crystal and that the possibility of reflection back into the metal should be allowed for at the boundary. This affects only the constant factor before the exponential, and this factor is very difficult to determine precisely. With this possible correction the equation becomes

$$I = AT^2 e^{-\frac{e\varphi}{kT}}. \tag{6.15}$$

where A is written for the product of a transmission factor and the numerical constant which is equal to 120 amp. per square centimeter per degree squared.

Equation (6.15) has been verified over as wide a range of the variable i as any equation in the subject of electricity; in this respect its generality is comparable to Ohm's law for metals. A schematic apparatus for testing it is shown in Fig. 6.15(a). The temperature of the cathode C is controlled by an internal insulated heating element. The value of the temperature may be measured with a thermocouple or optical pyrometer. The battery potential is high enough so that practically all of the electrons that evaporate from the cathode are drawn over to the anode or plate and contribute to the current measured by the milliammeter MA. The region surrounding the two electrodes is, of course, highly evacuated; such a two-electrode vacuum tube is known as a *diode.* A plot of the $\log_e$ of the current as a function of $1/T$ is approximately a straight line with a slope of $-e\varphi/k$. For accurate measurements extending over a large temperature range the T^2 factor of Eq. (6.15) must be considered and $\log_e i - 2 \log_e T$ must be plotted as a function of $1/T$ to obtain a straight line. It was in this way that the correctness of the T^2 factor was verified. The quantity φ, which is determined from the slope of this type of plot, is a function of the emitting surface. It is generally measured in volts and varies from about 1 for certain special surfaces to over 5 for clean nickel and platinum. As may be seen from Eq. (6.15), the lower the value of φ, the larger the thermionic current for a given temper-

ature. The constant A is also a function of the surface, though for most ordinary surfaces it is of the order of 60 amp. per square centimeter per degree squared. These constants are listed in Table IV for a number of representative surfaces.[1]

TABLE IV

Surface	A,* amp./cm.^{2} deg.^{2}	φ, volts
Barium	~ 60	2.11
Cesium	~ 60	1.81
Copper	~ 60	4.33
Gold	~ 40	4.90
Hafnium	~ 15	3.53
Molybdenum	~ 60	4.15
Nickel	~ 30	5.03
Palladium	~ 60	4.99
Platinum	~ 30	5.32
Silver	.....	4.74
Tantalum	~ 50	4.10
Thorium	~ 60	3.38
Tungsten	~ 60	4.53
Zirconium	~300	4.1
Thorium on tungsten (monatomic layer), approx.		2.7
Cesium on oxygen on tungsten, approx.		1.0

* This constant is a very sensitive function of surface conditions and the higher values listed are possibly unreliable.

6.6. Photoelectric Effects. *Photoconductive Effect.*—Certain semiconductors, of which selenium is a notable example, acquire a much higher conductivity when illuminated. This is readily understood in terms of the band theory of solids outlined in Sec. 6.3. The phenomena of the absorption and emission of light by the electronic structure of atomic systems require the concept that radiant energy is quantized and occurs in discrete amounts just as the energy levels of the electrons are limited to certain specified values. The energy unit for a given frequency ν of light is $h\nu$, where h is Planck's constant of Sec. 6.5. Such a unit of radiant energy is called a *photon*. When photons are incident on a selenium surface, they may be absorbed by electrons resulting in the elevation of these electrons in the band structure and the consequent enhancement of conductivity. There is a threshold frequency below which this effect is undetectable; and as the frequency of the incident light is increased, the increase in conductivity passes through a maximum for frequencies of the order of 1.5 to 2 times the threshold frequency. As

[1] References: DUSHMAN, *Rev. Mod. Phys.*, **2**, 381 (1930); REIMANN, "Thermionic Emission," John Wiley & Sons, Inc., New York, 1933; BECKER, *Rev. Mod. Phys.*, **7**, 95 (1934).

the mechanical properties of selenium are poor, a selenium cell is generally made by evaporating a thin film on a glass or porcelain surface upon which a double grid or set of interlocking combs of gold or platinum has previously been deposited to serve as electrodes (Fig. 6.16*a*). The range of wave length of light over which the conductivity is markedly affected is from about 6 to $9\mu (1\mu = 10^{-6}$ m.); the maximum response is at about 7μ. For a typical cell a change in illumination from 2 to 12 ft.-candles will decrease the resistance from 4 to 2 megohms. From these figures it

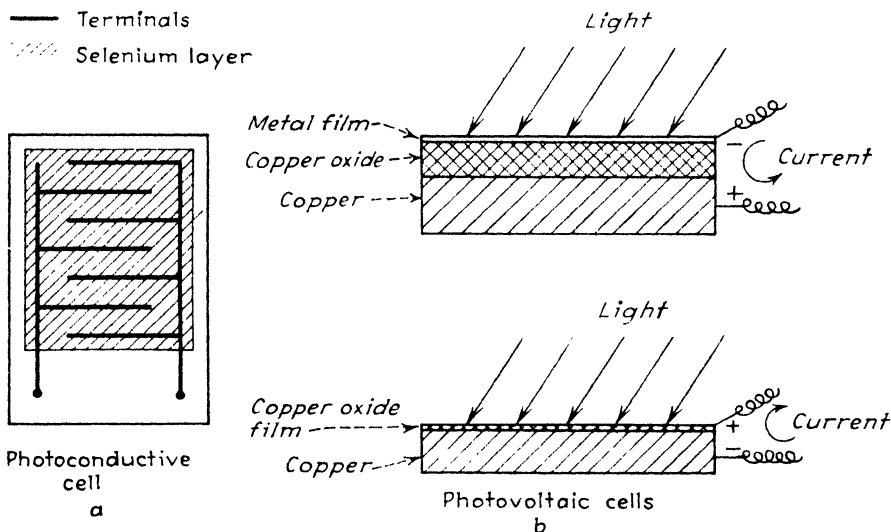



FIG. 6.16. Photoconductive and photovoltaic cells.

is seen that such a cell is a high-resistance device and should be matched by a high-resistance detecting or measuring circuit.[1]

[1] References: NIX, *Rev. Mod. Phys.*, **4**, 723 (1932); HUGHES, *Rev. Mod. Phys.*, **8**, 294 (1936).

Note on Photometric Units.—For the scientific specification of a source of radiation the radiant energy per unit solid angle per unit wave-length range must be given over the pertinent range of these variables. Units of these dimensions are used in pure scientific work. In much commercial work and for practical illumination measurement, an unrelated set of subjective units has been introduced. The unit of *luminous intensity*, the *standard candle*, is defined in terms of the brightness of radiation emerging from an orifice in the walls of an enclosure maintained at the temperature of freezing platinum. The brightness of this source is defined as 60 candles per square centimeter of opening. The brightness of different colors is determined by using certain standard luminosity factors. A source is said to be of *n candle power* if it is visually equivalent to *n* standard candles. Such equivalence is estimated by eye with a photometer. One candle is considered to emit 4π *lumens* of light flux in all directions, *i.e.*, 1 lumen per unit solid angle. The light flux per unit area, called the *illumination*, drops off as the inverse second power of the distance from the source. The unit is the lumen per square foot or *foot candle;* this is the illumination 1 foot from a

Photovoltaic Effect.—When the boundary between a metal and a semiconductor is illuminated, an emf. frequently develops between them. This effect is closely related to the rectifying properties of such boundaries as discussed in Sec. 6.3. The photon energy absorbed by electrons raise them in the energy-level structure, altering the mean electron energy in the same way that an applied emf. would. The transport of the electrons across the boundary is similarly affected differentially. Such boundaries are known as *photovoltaic junctions*. Two instances involving the semiconductor copper oxide are shown in Fig. 6.16*b*. In the upper figure the boundary influenced by the light is that between a transparent metal film and the underlying oxide. As there is generally excess oxygen in the oxide lattice at such a boundary and oxygen atoms form local acceptor levels for electrons, the junction is that between a P-type semiconductor and a metal; electrons thus tend to move into the metal under the influence of light, and it becomes negative as shown. In the lower figure the oxide itself forms the transparent layer, and the action takes place at the boundary with the copper. Here there are excess copper atoms in the oxide lattice, the semiconductor is of the donator or N type, and electrons tend to move from the metal into the semiconductor under the influence of light.

The characteristics of a commercial cell are shown in Fig. 6.17. The cell emf. in terms of the luminous flux striking the surface is given by the dashed curve. The resistance of the cell also changes with illumination in such a way that the current established in a low-resistance external

---

standard candle. A practical source does not emit light equally in all directions so the direction of observation must be specified. A source of n candle power in an indicated direction emits n lumens per unit solid angle in that direction and at a distance of f ft. the illumination is n/f^2 foot-candles. The practical comparison standard is generally a calibrated incandescent lamp. An idea of the efficiency of conversion of electric energy into light in various types of lamps is given by the following set of approximate figures:

Type of lamp	Efficiency	
	Candle power/watt	Lumens/watt
Carbon-filament incandescent lamp	0.2	2.5
Tungsten filament in high vacuum	0.6	7.5
Tungsten filament in an inert gas	1.6	20
Quartz-mercury arc	4.0	50 (and greater)

Photometric units are related approximately to the more usual scientific units through the fact that 1 watt of radiant flux in the wave-length region of maximum visibility (circa 0.57μ) is equivalent to 668 lumens.

circuit is very closely proportional to the illumination of the cell. Thus a combination of this cell and a microammeter is particularly well suited for illumination measurements, photographic-exposure meters, etc. The solid lines give the circuit current for two values of the external resistance as a function of the illumination. An ordinary incandescent lamp with the filament at about 3000°C. was used to obtain these curves. The spectral response of the cell is very similar to the normal visibility curve shown in Fig. 6.22. This is an added advantage for this type of cell in

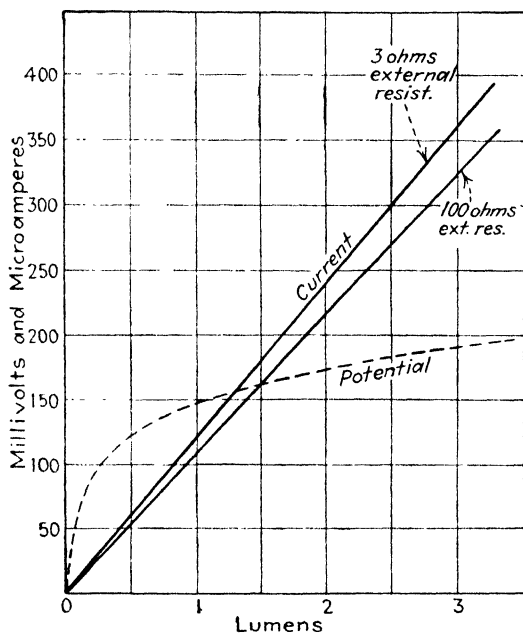



FIG. 6.17.—Curves representing the performance of the Weston photronic cell.

illumination measurement as the quantitative measurements made with it agree closely with visual estimates for a wide variety of sources.[1]

Surface Photoelectric Effect.—The term surface photoelectric effect, or merely photoelectric effect, refers to the liberation of electrons from a surface under the influence of light. It was the discovery and investigation of this phenomenon early in the present century that lead to the photon conception of radiation. The fundamental experimental facts may be elicited by means of an apparatus of the type shown diagrammatically in Fig. 6.18. *a* and *b* are two metal electrodes in an evacuated chamber. The potential difference between them measured by the voltmeter *V* is variable and a galvanometer *G* records the current flowing in the circuit. If the plates are shielded from light, there is no current across the evacuated region. If plate *a*, for instance, is illuminated and

[1] Reference: GRONDAHL, *Rev. Mod. Phys.*, **5**, 141 (1933).

the battery is applied in the proper sense, a current may be detected in the circuit corresponding to a flow of electrons from *a* to *b*. The magnitude of this current is proportional to the intensity of illumination and there is no lag (less than 10^{-8} sec.) between the illumination and the inception of the current.

The most important information concerning this effect comes from an analysis of the current as a function of the retarding potential necessary to reduce it to zero and the wave length or frequency of the light falling on the plate. For light of any one frequency the characteristic curve of current as a function of potential resembles the high-vacuum curve of Fig. 6.21. The saturation current at large accelerating voltages depends on the intensity of illumination. The value of the voltage at which the current curve drops to zero is independent of the amount of illumination, but it is a function of the surface *b* and the frequency of the light. This voltage necessary to reduce the current to zero actually corresponds to a retarding field against which the electrons must move to reach the plate *b*. If now this critical cut-off voltage, which is a function of the wave length of the monochromatic light striking the plate, is plotted as a function of the frequency of the light, the straight line shown in Fig. 6.19 is obtained. The X's indicate experimental points obtained, for instance, by using selected lines from the mercury spectrum.

FIG. 6.18.—V to V_m as shown.

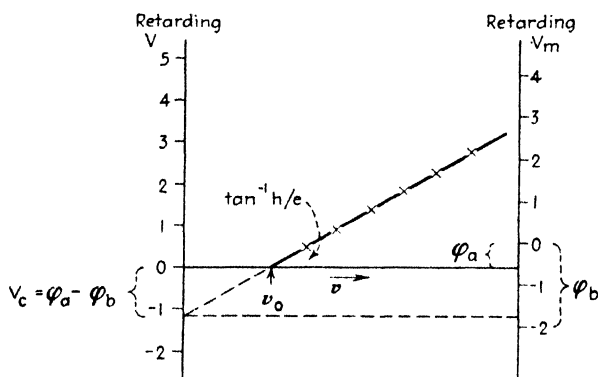



FIG. 6.19.—Analysis of the photoelectic equation.

This linear relation means that the energy possessed by a photoelectron, which enables it to move against the retarding field from plate *a* to plate *b*, is proportional to the frequency of the light. These phenomena can be satisfactorily accounted for only on a theory which considers a beam of light to be a stream of energy units or photons, as described

earlier in this section. Aside from its practical uses the photoelectric effect is most important as providing convincing evidence for this point of view. The considerations which determine the trajectories of the photons will be discussed later in connection with the electromagnetic theory of radiation.

A conduction electron at the surface of the metal absorbs the energy of a photon and is thus enabled to escape against the surface forces. The energy necessary to cross the surface is $e\varphi_a$, where φ_a is the work function of surface a and the additional energy necessary to reach plate b against the retarding potential V, which is assumed to exist between the plates, is eV. This total energy must have been supplied by the photon

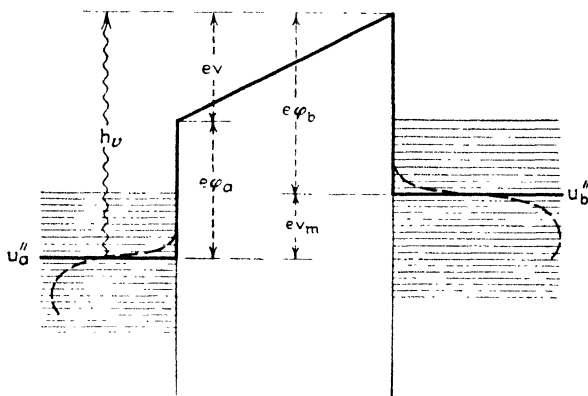



FIG. 6.20.—Schematic representation of the energy relations involved in photoelectric emission.

energy $h\nu$. Thus the energy equation, which is known as Einstein's equation, may be written

$$h\nu = e\varphi_a + eV \tag{6.16}$$

The true potential difference between the surfaces is not the meter reading V_m, which represents the difference in potential between the Fermi levels u'' of the two metals, but a study of Fig. 6.20 shows that

$$V = V_m - (\varphi_a - \varphi_b)$$

Therefore

$$V_m = \frac{h}{e}\nu - \varphi_b \tag{6.17}$$

The slope of the straight line of Fig. 6.19 which is $dV_m/d\nu$ is thus h/e. With e known from Millikan's experiment the fundamental constant h can be determined from this slope. It is found to be 6.62×10^{-27} erg-sec. or 6.62×10^{-34} joule-sec. If ν is less than a critical value ν_0, given by $e\varphi_a = h\nu_0$, the electrons do not gain sufficient energy to escape from the metal. This is known as the threshold frequency or the corresponding

wave length as the long-wave-length limit. V_m corresponding to this limit is written as V_c and is known as the *contact potential* difference between the surfaces. The equation is not applicable for smaller values of ν, but extrapolation to $\nu = 0$ yields

$$V_m = V_c - \varphi_a = -\varphi_b.$$

These various quantities are indicated in Fig. 6.19. To make a successful measurement of the constant h by an experiment of this type, φ_b must be sufficiently greater than φ_a, and the frequencies used must lie in the range

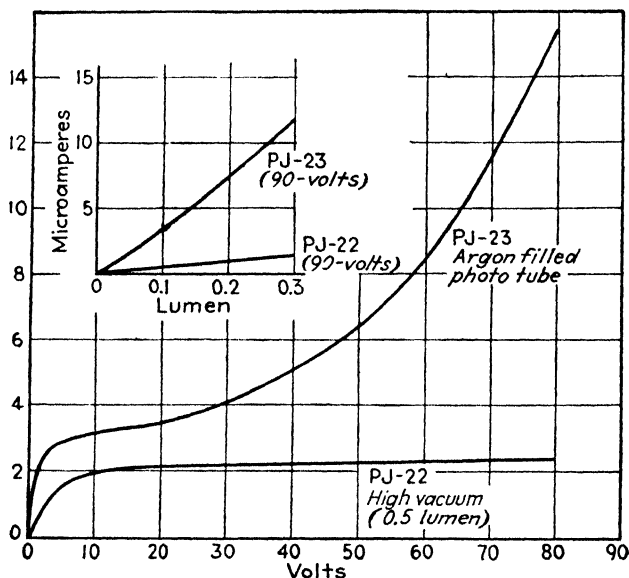



FIG. 6.21.—Representative phototube characteristics.

$e(\varphi_b - \varphi_a)/h$, so that no electrons are liberated from plate b as these would confuse the results.

Commercial photoelectric cells are both of this high-vacuum type and of a gas-filled type. An inert gas such as argon is generally used in the latter. Its advantage is that ionization of the gas permits larger currents to flow for the same amount of illumination. However, care must be used not to apply too great a voltage to this type of cell or the active surface may be impaired; also, the response is not as rapid as for the high-vacuum tube. Typical characteristics of high-vacuum and gas-filled cells are shown in Fig. 6.21. They are seen to be nonlinear and may be used for any of the purposes for which a nonlinear element is suited. The illumination response, however, is seen to be linear and of the same order as for other types of photocells. The wave-length range over which the cells respond is determined by the nature of the cathode

surface and the transmission characteristics of the glass envelope. A quartz envelope may be used for the ultraviolet region of the spectrum. Figure 6.22 indicates the response characteristics for an ordinary potassium surface and for a carefully prepared surface of cesium on oxidized silver. The latter is seen to extend well out into the infrared. The ordinary visibility curve and the energy-distribution curve for a typical incandescent lamp are included for comparison.

As may be seen from their characteristics, the internal resistance of these cells is high and they are well suited to a high-resistance amplifying device such as a thermionic tube. Their numerous applications all depend on the possibility of such amplification. Suitable circuits

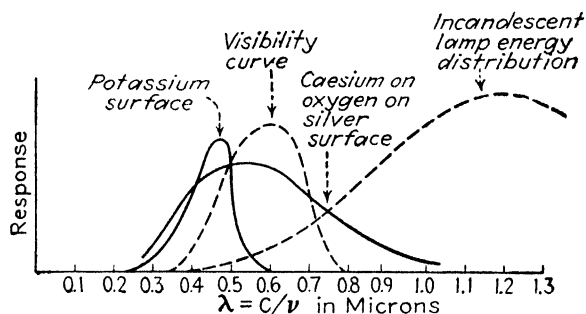



FIG. 6.22.—Spectral response of representative photoelectric surfaces.

will be indicated in a later section. With sufficient amplification they may be used for controlling large amounts of power, and their applications are almost innumerable. Anything which can intercept or alter the character of a beam of light can be made to perform a desired mechanical operation. The presence of a person may cause a door to open or an alarm to ring. The number of people passing a given point or only those passing in a specified direction may be counted. Many other counting, sorting, and color-judging processes may be performed. The decrease in illumination caused by the presence of smoke or dust particles in the air may be made to operate a control device. These cells are also widely used for sound reproduction in the moving-picture industry. The sound track on a film is made up of regions of variable photographic density whose pattern corresponds in frequency and intensity to the original sound. This intercepts a narrow beam of light directed on a photoelectric cell, and the cell current reproduces the film pattern and is amplified and delivered to a loud-speaker. A linear response to radiation is, of course, desirable for this type of reproduction work. Though the gas-filled cell is less linear than the high-vacuum type, it is quite adequate for sound work.[1]

[1] ZWORYKIN and WILSON, "Photocells and Their Application," John Wiley & Sons, Inc., New York (1930); HUGHES and DUBRIDGE, "Photoelectric Phenomena,"

Secondary Emission.—It should be mentioned that electrons in a metal can gain sufficient energy to escape from the surface not only from thermal energy and incident photons but also from electrons, ions, or atoms striking the surface if these possess sufficient energy. Incident electrons are more efficient in liberating other electrons than are the heavier atoms or ions, but both play roles in practical instances discussed in subsequent chapters (Secs. 7.3 and 8.1). The phenomenon is known as *secondary emission.* If the electrons bombarding a typical surface have energies in excess of about 10 electron volts, an appreciable fraction of them will give rise to secondary electrons. For special surfaces the number of secondaries per primary may rise to 5 or 10 at bombarding energies of the order of 100 electron volts. This phenomenon is used to provide amplication for very feeble electron currents. By the suitable arrangement of electric or magnetic fields the electrons leaving a surface that can be called the cathode can be directed against an anode surface specially prepared to enhance secondary emission and the secondaries from this surface in turn directed against a second anode, etc., through a series of eight or ten plates. At each plate the electron current will be increased by a factor f, so that after n anode surfaces the original cathode current will be greater by the factor f^n. Taking a typical value of f as 5 and assuming n to be 10, the over-all multiplication is seen to be about 10^7. Such devices are of great value in atomic research where small electron currents are involved, and they are also of commercial importance as very high-gain amplifiers. Their virtue lies not only in the very high amplification but also in the more favorable signal-to-noise ratio that can be obtained with them.

Problems

1. A mass of 1.7396 gm. of copper is deposited by a constant current flowing through a copper voltameter for 1.5 hr. An ammeter in the circuit reads 1 amp. during the process. Calculate the correct value of the current and the error of the meter.

2. How many atoms of silver are deposited per second by a current of 1 ma. flowing through a silver voltameter? What mass of silver is deposited in 30 minutes?

3. How long will it take to fill a balloon 1 meter in diameter to a pressure of 1.2 atmospheres of hydrogen by electrolysis at a current of 10 amperes?

4. From the value of the faraday and the electronic charge calculate the number of atoms in a mole of a substance. How many molecules are there in a cubic centimeter (*a*) of water, (*b*) of air (density 0.0013 gm. per cubic centimeter)?

5. Calculate the electrical resistance of a 50 per cent $AgNO_3$ solution contained between two metal plates 20 cm. on a side and 2 cm. apart.

6. An electrolytic cell is composed of a glass-bottomed cylindrical metal container 20 cm. high and 5 cm. in diameter filled with a 10 per cent solution of NaCl; a central cylindrical rod 1 cm. in diameter extending down to the glass bottom forms the other electrode. From Sec. 3.4 and Table I calculate the resistance of the cell.

7. An electrolytic cell is found to have a resistance of 100 ohms when measured in a bridge at a very high frequency. At a frequency of 100 cycles per second there is a phase lag of 0.1 rad. between the current and the voltage applied to the cell. Calculate the polarization constant P and the effective resistance of the cell at 100 cycles per second.

8. Two platinum electrodes, one in a porous cup containing a 50 per cent HNO_3 solution and the other outside the cup in a 10 per cent solution of the acid, form a

---

McGraw-Hill Book Company, Inc., New York, 1932; LINFORD, *Rev. Mod. Phys.*, **5**, 34 (1933).

concentration cell. Assuming the mobility of the positive ion is three times that of the negative one, calculate the emf. of the cell.

9. A voltage cell composed of a zinc cathode, a zinc sulphate solution, and a mercury anode (Clark cell) has an emf. of 1.433 volts at 15°C. and a temperature coefficient of 0.0012 volt per degree Centigrade. Calculate the heat of reaction of the cell process at 15°C.

10. Two semicircular rings, one of copper and one of nickel, are joined together to form a ring 10 cm. in diameter with a cross section of 2 cm.² If one junction is kept at 0°C. and the other at 100°C., calculate the current that will circulate in the ring (neglect the junction resistances).

11. The cold junction of a copper-constantan thermocouple is maintained at 0°C. Plot the emf. developed as a function of the temperature of the hot junction from 0 to 500°C.

12. For what temperature difference between the junctions of a copper-iron thermocouple is the emf. developed, (*a*) a maximum, (*b*) zero?

13. Calculate the Peltier coefficient for a copper-nickel junction at 0°C. and also calculate the values of the Thomson coefficients for the two metals at that temperature.

14. The emf. developed by a chromel-alumel thermocouple, the cold junction of which is at 0°C., is 18.37 mv. and 44.24 mv. for the hot junction at the boiling point of sulphur and the melting point of copper, respectively. Calculate the constants of Eq. (6.11) for these two substances.

15. Plot the electrostatic potential energy of an electron as a function of its distance above the supposedly plane surface of a metal on the assumption that $\varphi_0 = 3$ volts and that an extracting field of 10^4 volts per centimeter is applied to the surface. At what distance out will the total force be such as to pull an electron away from the surface?

16. What thermionic current may be drawn from a plate of tungsten 1 cm.² in area which is maintained at a temperature of 3300°C. (*a*) for a negligible surface field, (*b*) for an extracting field of 10^7 volts per meter? (Assume that $A = 60$ amp. per square centimeter per degree squared.)

17. Calculate the thermionic current that can be drawn from a tantalum filament 0.1 mm. in diameter and 10 cm. long at a temperature of 3000° abs. assuming a negligible surface field.

18. A photocell whose characteristics are given by Fig. 6.17 has an area of 1.7 in.² and is placed 10 ft. from a 150-candle-power lamp. What is the emf. developed?

19. The cell of the previous problem, when placed 12 ft. from a lamp, sends a current of 300 μa. through a meter with a resistance of 3 ohms. What is the candle power of the lamp?

20. When light from a mercury arc with the wave length 0.5461 μ shines on the cathode of a photocell, a retarding potential of 1.563 volts (as read by a potentiometer) must be applied to suppress the photoelectric current. For light of wave length 0.4047 μ a retarding potential of 2.356 volts must be applied to accomplish the same thing. Calculate the value of Planck's constant h, assuming Millikan's value for e. If the long-wave-length limit of the surface is 0.6500 μ, calculate the work functions of the electrodes and the contact potential between them.

21. Calculate graphically from Fig. 6.21 for both types of cell the potential difference that will be developed across a resistance of 2 megohms in series with the cell and a 90-volt battery in the presence of 0.5 lumen. Plot the variation of this potential difference as the potential of the battery is varied from 0 to 90 volts for the two types of cell.

22. See how well the electrolytes listed in Table I satisfy the Debye-Hückel equation for strong electrolytes given in Sec. 6.1.

23. Photoelectric electrons leave the surface of a cathode under the influence of light of frequency ν. They are accelerated across an intervening space toward an anode in the shape of a box with a small opening facing the cathode. Show that those electrons which enter the box have inside it a kinetic energy given by $h\nu + e(V_n - \varphi_b)$, where V_n is the potential difference between the anode and cathode as measured by a voltmeter and φ_b is the work function of the anode.

CHAPTER VII

THERMIONIC VACUUM TUBES

7.1. The Diode.—In Sec. 6.5 it was seen that electrons are liberated from a conducting surface if the latter is maintained at a sufficiently high temperature. If this hot surface is in the presence of another conducting surface and the terminals of a battery are connected between them in such a way that the hot surface is the cathode (−) and the other surface the anode (+), a current of electrons will flow through the intervening space from the cathode to the anode. If the region surrounding the electrodes is highly evacuated, there are no gas molecules present to interfere with or otherwise influence the electron flow. Such a device is known as a high-vacuum thermionic tube, or simply as a vacuum tube. The presence of a gas between the electrodes profoundly modifies the current flow; and a consideration of gas-filled tubes will be deferred till the following chapter.

The simplest type of thermionic vacuum tube is known as the *diode* or *kenotron*. It consists simply of a filament or indirectly heated cathode which emits electrons and a positive plate or anode to which they are drawn. These are, of course, enclosed in a highly evacuated envelope. Glass is generally used for this envelope, though it can be made of metal if glass seals are used for the admission of electrode leads. The most important uses of these tubes are for rectification and demodulation, though the Coolidge X-ray tube for the production of very high frequency electromagnetic radiation is also of this type.

Figure 6.15 illustrates schematically this type of tube. The potential distribution between the cathode and anode cannot be calculated simply from the geometry of the electrodes. The cloud of electrons carrying the current through the intervening space has a pronounced effect on this potential distribution. In terms of lines of force many of those leaving the anode terminate on electrons in the space and in most geometrical arrangements certain of the lines from the cathode also terminate on electrons; hence there is a field retarding the motion of electrons from the cathode, and a potential minimum exists somewhere in the space between. The potential distribution will be calculated from Poisson's equation for the case of plane electrodes, with the simplifying assumption that the

electrons start from rest at the cathode (the potential minimum is then at the surface of this electrode). Assuming the electrodes to be normal to the x axis, Poisson's equation (Sec. 1.7) reduces to

$$\frac{d^2V}{dx^2} = -\frac{q_v}{\kappa_0} \tag{7.1}$$

q_v can be expressed in terms of V from the following considerations. The current per unit area i is equal to $-q_v u$ where u is the electron velocity at a distance x from the cathode. If the zero of potential is chosen at the cathode and if there are no gas molecules in the space to which the electrons can transfer their energy, the kinetic energy of an electron must be equal to the electrical energy gained from the field, *i.e.*, $\frac{1}{2}mu^2 = eV$. The above equation then reduces to:

$$\frac{d^2V}{dx^2} = \frac{i}{u\kappa_0} = \frac{i}{\kappa_0}\left(\frac{m}{2e}\right)^{1/2} V^{-1/2}$$

This may be integrated by multiplying the left side by $2(dV/dx)dx$ and the right by its equivalent $2dV$. The result is readily seen to be

$$\frac{dV}{dx} = 2\left(\frac{i}{\kappa_0}\right)^{1/2}\left(\frac{mV}{2e}\right)^{1/4} \qquad (= -\mathbf{E}) \tag{7.2}$$

if the field is assumed to be zero at the cathode which makes the constant of integration vanish. The variables are separable so the equation can be immediately integrated a second time to yield

$$V = \left(\frac{9i}{4\kappa_0}\right)^{2/3}\left(\frac{m}{2e}\right)^{1/3} x^{4/3} \tag{7.3}$$

The constant of integration again vanishes from the assumption that the zero of potential is at the cathode surface. Figure 7.1 is a plot of the three quantities V, $-E$, and $-(9/40\kappa_0)q_v$ from Eqs. (7.1), (7.2), and (7.3) for a unit distance between cathode and anode and a current density of $4\kappa_0/9\sqrt{2e/m}$. The electron density falls rapidly from cathode to anode, the absolute value of the field and the potential both increase.

If the anode is at a distance d from the cathode and V is its potential, Eq. (7.3) represents the characteristic of the device considered as a circuit element. Setting $x = d$ and rearranging

$$i = \frac{4\kappa_0}{9}\sqrt{\frac{2e}{m}}\frac{V^{3/2}}{d^2} \tag{7.4}$$

Thus the current is proportional to $V^{3/2}$ and inversely proportional to the square of the electrode separation. It should be emphasized that this assumes an unlimited supply of electrons at the cathode.

Actually, of course, this is not the case for the supply of electrons is limited by the cathode temperature. At relatively low potentials where Eq. (7.4) applies the current is said to be *space-charge-limited.* At considerably higher potentials the saturation current which is determined by the cathode temperature is drawn. Typical characteristics

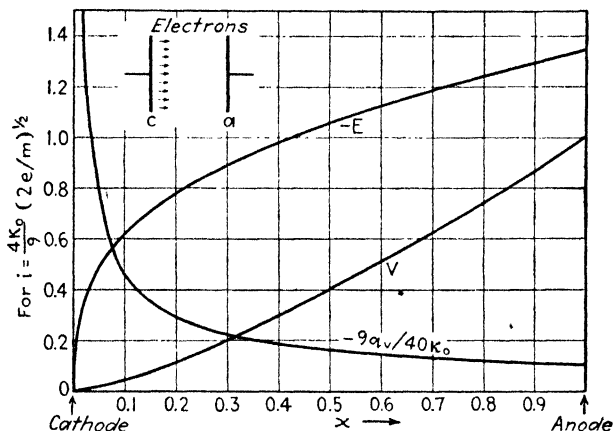



FIG. 7.1.—Potential, field, and charge distribution for a space-charge-limited current between plane parallel electrodes.

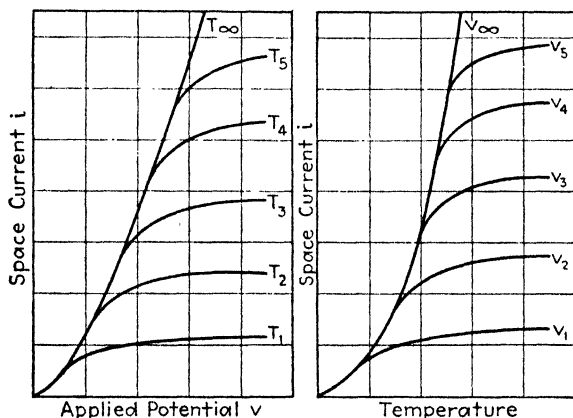



FIG. 7.2.—Variation of thermionic current with applied potential and temperature.

for various temperatures are illustrated at the left in Fig. 7.2. At the right are shown a series of i-T curves for various anode potentials. A similar calculation can be performed for a cylindrical cathode of radius a surrounded by a cylindrical anode of radius b. The space current per unit length of the cylinders is then given by

$$i = \frac{8\pi\kappa_0}{9}\sqrt{\frac{2e}{m}}\frac{V^{3/2}}{b\beta^2} \tag{7.5}$$

where $\beta \cong \log_e (b/a) - \frac{2}{5}[\log_e (b/a)]^2 + \frac{11}{120}[\log_e (b/a)]^3$, which for large b/a reduces to unity. Here again it is seen that the current is proportional to the $\frac{3}{2}$ power of the anode potential and, in fact, it can be shown perfectly generally that this is the case for any geometrical arrangement. An assumption in the derivation of the $\frac{3}{2}$ power relationship was that the cathode constituted an equipotential surface. If the cathode is heated directly by carrying a current, the difference in potential between its two ends causes deviations from this simple relationship. If the anode is returned to the negative end of the cathode, the anode current is reduced by this effect; if returned to the positive end, the current is greater. For filaments heated by alternating currents this can lead to "hum," which in practice is greatly reduced by returning the plate to the center point of the cathode.

When used simply as a nonlinear element, the type of nonlinearity desired determines the electrode spacing, cathode temperature, etc. For rectification there should be a minimum of power dissipation in the rectifier for a given current. From the above equations this is seen to imply a close spacing of the electrodes. Cathode-heating power must also be considered and to reduce this, heat shields and cathode surfaces of low work function are used where possible. In the case of the X-ray tube, the tube itself is the load and a maximum of power is to be dissipated in it. The applied potential is very high and the current is temperature-limited.

7.2. The Triode.—In the diode the space current is limited by the electron cloud which extends with decreasing density from cathode to anode and alters the potential distribution in this region. If a third electrode is interposed between these two, the potential distribution can be further altered and the flow of space current controlled. This electrode generally takes the form of a coarse mesh grid or wire helix surrounding the cathode and is known as the *grid*. A tube with such a control grid is called a *triode* or *pliotron*. It is a special type of three-terminal device in which electrons can flow between cathode and grid and cathode and plate. In the absence of complicating factors such as gas or secondary emission there is no current flow between grid and plate. Capacitative interaction of these two electrodes will be discussed later. As the tube is generally used, the grid is kept negative with respect to the cathode, so the only current flowing is that to the plate; but this is controlled both by the potential of the grid and that of the plate itself. While the general shape of the characteristics of a triode resembles that of a three-halves power curve, it is not convenient for practical purposes to analyze the behavior of a triode on this basis. The Taylor's series method is both more widely applicable and more instructive; any desired degree of approximation can be obtained, though only the linear and quadratic ones find much application.

Retaining for the moment the possibility of a grid current i_c, both it and the plate current i_b are functions of the grid and plate potentials

e_c and e_b, respectively. These functions, $i_c(e_c, e_b)$ and $i_b(e_c, e_b)$ can be expanded in a Taylor's series (Appendix A)

$$i_c = I_c + \underbrace{\left(\frac{\partial i_c}{\partial e_c}\right)e_g + \left(\frac{\partial i_c}{\partial e_b}\right)e_p}_{\text{linear approx.}} + \underbrace{\frac{1}{2}\left\{\frac{\partial^2 i_c}{\partial e_c^2}e_g^2 + 2\frac{\partial^2 i_c}{\partial e_c \partial e_b}e_g e_p + \frac{\partial^2 i_c}{\partial e_b^2}e_p^2\right\}}_{\text{quadratic approx.}} + \cdots$$

$$i_b = I_b + \underbrace{\left(\frac{\partial i_b}{\partial e_c}\right)e_g + \left(\frac{\partial i_b}{\partial e_b}\right)e_p}_{\text{linear approx.}} + \underbrace{\frac{1}{2}\left\{\frac{\partial^2 i_b}{\partial e_c^2}e_g^2 + 2\frac{\partial^2 i_b}{\partial e_c \partial e_b}e_g e_p + \frac{\partial^2 i_b}{\partial e_b^2}e_p^2\right\}}_{\text{quadratic approx.}} + \cdots$$

Here e_g and e_p are the variations (generally small) of the independent variables e_c and e_b about the points of expansion and I_c and I_b are the so-called quiescent values of i_c and i_b, *i.e.*, the values of these currents when $e_g = e_p = 0$. To proceed with approximations beyond the linear one would lead too far for an elementary treatment. The general phenomena involved have already been discussed in Chap. V. From time to time reference will be made to the higher approximations, but the present chapter will be largely confined to the linear one. This implies, of course, that the partial differential coefficients are considered to be constant, *i.e.*, the characteristics are straight lines, or the independent potential variables are only subject to infinitesimal changes.

The partial differential coefficients are so useful that they have received special names and abbreviations. By analogy with Sec. 5.1

$$\left(\frac{\partial i_c}{\partial e_c}\right) = k_g = \frac{1}{r_g} \text{ (grid conductance)} \left(\frac{\partial i_b}{\partial e_b}\right) = k_p = \frac{1}{r_p} \text{ (plate conductance)}$$

where r_g and r_p are the dynamic grid and plate resistances, respectively. Similarly

$$\left(\frac{\partial i_c}{\partial e_b}\right) = s_g \text{ (reflex transconductance)} \quad \left(\frac{\partial i_b}{\partial e_c}\right) = s_p \text{ (transconductance).}$$

With these abbreviations and writing $i_c - I_c = i_g$ and $i_b - I_b = i_p$ for the variations in the dependent currents, the linear approximations become

$$i_g = k_g e_g + s_g e_p \tag{7.6}$$
$$i_p = s_p e_g + k_p e_p \tag{7.6'}$$

Similarly, of course, e_b, for instance, could be considered to be a function of e_c and i_b leading to the linear terms in the Taylor's series

$$e_b = E_b + \left(\frac{\partial e_b}{\partial e_c}\right)e_g + \left(\frac{\partial e_b}{\partial i_b}\right)i_p$$

where E_b is the quiescent plate potential. With the abbreviation $(\partial e_b/\partial e_c) = -\mu_p$ (*plate amplification factor* frequently written μ), and as

$e_b - E_p = e_p$, this equation may be written

$$e_p = -\mu_p e_g + r_p i_p \qquad (7.7)$$

Similarly, considering e_c a function of e_b and i_c and introducing the grid amplification factor $-\mu_g = (\partial e_c / \partial e_b)$ ($\mu_g \neq 1/\mu_p$ as the current variable involved is different in the two cases)

FIG. 7.3.—The triode.

$$e_g = r_g i_g - \mu_g e_p \qquad (7.7')$$

There are certain useful relations between these coefficients. Eliminating i_p between Eqs. (7.6) and (7.7)

$$e_p = (-\mu_p + r_p s_p)e_g + r_p k_p e_p$$

Since e_p and e_g are independent, the coefficient of e_g must be zero, and that of e_p on the right is unity. Hence

$$\mu_p = r_p s_p \qquad (7.8)$$

Similarly

$$\mu_g = r_g s_g \qquad (7.8')$$

Referring to Fig. (7.4), Eqs. (7.6) and (7.6') can be written in various forms. E_B is the potential of the plate battery and E_{Bg} that of the grid battery; e_0 is a small variable potential in the grid circuit. Consider first the plate circuit.

$$E_B = e_b + e_d = E_b + E_d + e_p + e_l$$

where E_d and e_l are the constant and variable components of the load potential. When

$$e_p = 0, \qquad E_B = E_b + E_d \qquad (7.9)$$

therefore in general

$$e_p = -e_l = -r_l i_p$$

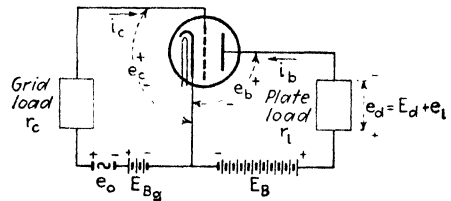



FIG. 7.4.—Triode circuit.

Expressing e_p in Eq. (7.6') in terms of i_p by this relation

$$i_p = \frac{s_p}{1 + r_l k_p} e_g$$

And by Eq. (7.8)

$$i_p = \frac{\mu_p}{r_p + r_l} e_g \qquad (7.10)$$

Or expressing the same equation in terms of e_l and e_g

$$e_l = \frac{s_p}{k_p + k_l} e_g \qquad (7.11)$$

where k_l is the load conductance. Equations (7.10) and (7.11) are two forms of the equivalent plate-circuit (e.p.c.) theorem. A similar theorem can be derived for the grid circuit. Since $e_0 = e_g + r_c i_g$, Eq. (7.6′) becomes

$$i_g = \frac{\mu_g e_p + e_0}{r_g + r_c} \tag{7.10′}$$

Since the grid current is generally negligible, this equation which represents the equivalent grid-circuit (e.g.c.) theorem is not of so much importance as the previous ones.

FIG. 7.5.—Simple equivalent circuit of a triode.

These equivalent-circuit theorems are very useful in that they allow the grid and plate circuits to be considered separately. These separate circuits are shown schematically in Fig. 7.5; the fact that there is a certain connection between them is indicated by the common point. From Eq. (7.10′) the grid circuit contains resistances r_c and r_g with an emf. e_0 and an effective but ficticious emf. $\mu_g e_p$. From Eq. (7.10) the plate

FIG. 7.6.—Graphical analysis of the linear operation of a triode.

circuit is made up of resistances r_l and r_p with an effective but ficticious emf. $\mu_p e_g$. These emfs. are ficticious in the sense that in the absence of gas in the tube or secondary electron emission from either element or capacitative effects between them (all of which are neglected in the present account) there is no power interchange between the two circuits.

The graphical analysis of the plate circuit of a triode is particularly useful and instructive. The coordinates in Fig. 7.6 are the plate potential and plate current; to represent the grid potential as well would, of course, require three dimensions. Instead, the family of curves cor-

responding to a series of values of the grid potential are drawn upon this diagram. They may be thought of as the projections on this plane of the intersections of parallel planes, equally spaced along the e_c axis, and the plate-current surface $i_b(e_b, e_c)$. The plate load r_l is represented by the straight line extending back from the battery potential E_B and making an angle $\tan^{-1} r_l$ with the i_b axis. The quiescent point Q ($e_p = i_p = 0$) has the coordinates E_b, I_b. The heavy line extending on either side of Q from $e_c = -9$ ($e_g = 6$) to $e_c = -21$ ($e_g = -6$) represents the motion of the plate point for this change in grid potential. If a sinusoidal potential wave of this amplitude, for instance, is impressed upon the grid, the approximately sinusoidal variations in plate potential and plate current are shown by the curves on the auxiliary time axes above and to the right. If enough grid lines are drawn in from experimental data,

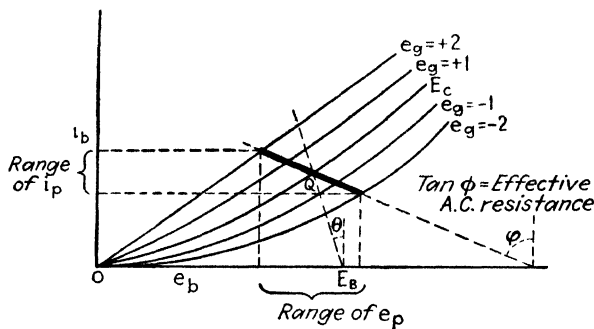



FIG. 7.7.—Load with different alternating- and direct-current resistances.

the variation in plate potential and current can be plotted point by point for any arbitrary form of grid excitation. $E_B - e_b$ is the potential drop in the load resistance and e_b is the drop in the tube. The parameters μ_p, s_p, and r_p can also be interpreted in terms of this diagram. If the linear approximation is good for an extended region around Q (the e_g lines are straight and parallel to one another), these parameters are constant in this region and may be represented by the lines ab and bc and their ratio. From the previous definitions

$$ab = \text{change in } e_b \text{ per unit change in } e_g \text{ at constant } i_b = \mu_p$$
$$bc = \text{change in } i_b \text{ per unit change in } e_g \text{ at constant } e_b = s_p$$
$$\frac{ab}{bc} = \text{change in } e_b \text{ per unit change in } i_b \text{ at constant } e_g = r_p$$

It is not necessary that the alternating- and direct-current resistance of the load should be the same for the linear approximation to hold quite adequately. The only additional requirement is that load characteristic should be approximately a straight line in the operating region. Figure 7.7 represents such a situation. The heavy line represents a straight

portion of the load characteristic on either side of the point Q. The direct-current load resistance is tan θ but the dynamic or alternating-current resistance is tan ϕ. The most common example of such a load is a tuned circuit at resonance; this will be more fully discussed in later sections where the advantages of such an arrangement will become evident.

The power relations in the plate circuit of a triode can also be considered either analytically or graphically. The power equation is of course

$$E_B i_b = e_b i_b + e_d i_b$$

These e's and i's can be expressed in terms of their quiescent and incremental values and integrated over a complete period, *e.g.*, one wave of Fig. 7.6. As the large letters represent constants and the average values of e_p, e_l, and i_p are zero the terms of the form

$$\frac{1}{\tau}\int_0^\tau E_B i_p\,dt, \qquad \frac{1}{\tau}\int_0^\tau E_b i_p\,dt, \qquad \frac{1}{\tau}\int_0^\tau e_p I_b\,dt, \qquad \frac{1}{\tau}\int_0^\tau E_d i_p\,dt, \qquad \frac{1}{\tau}\int_0^\tau e_l I_b\,dt$$

vanish. The average power equation then becomes

$$\underbrace{E_B I_b}_{P_B} = \underbrace{E_b I_b}_{P_b} + \underbrace{\frac{1}{\tau}\int_0^\tau e_p i_p\,dt}_{p_p} + \underbrace{E_d I_b}_{P_d} + \underbrace{\frac{1}{\tau}\int_0^\tau e_l i_p\,dt}_{p_l}$$

From Eq. (7.9)

$$P_B = P_b + P_d \qquad \text{and hence} \qquad p_p = -p_l$$

Thus, though the source of power is the plate battery, the tube acts as if it were a generator with an internal resistance r_p. p_l is generally positive and p_p negative, *i.e.*, power is supplied by the tube to the load. In the particular case of sinusoidal variation of e_l and i_p (sinusoidal variation of e_g in the linear region of the tube characteristic)

$$p_p = \frac{1}{\tau}\int_0^\tau e_p i_p\,dt = \frac{E_{pm} I_{pm}}{\tau}\int_0^\tau \sin^2 \omega t\,dt = \frac{E_{pm} I_{pm}}{2} = E_p I_p$$

where E_{pm} and I_{pm} are the maximum and E_p and I_p the effective values of the plate potential and current waves, respectively. These power relations are illustrated in Fig. 7.8. From the above equations

P_B = area of the rectangle *abcd*
P_b = area of the rectangle *abfe*
P_d = area of the rectangle *efcd*
$p_p = -p_l$ = area of the triangle *ehg*

The efficiency of the tube as a linear device for transforming battery power into alternating-current load power is a function of the maximum grid swing E_{gm} and is generally very low. It is the ratio of the areas $ehg/abcd$, and even in the most favorable case of linearity throughout the entire diagram (which cannot be realized) and $r_p = r_l$ the area ehg can only equal $abfe/2$, *i.e.*, only half the direct-current power supplied to the

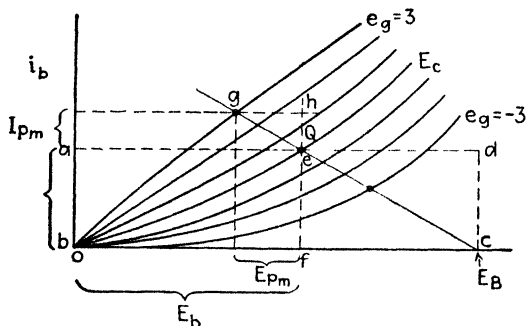



FIG. 7.8.—Power relations.

plate can be transferred as alternating-current power to the load. These considerations, of course, do not hold for nonlinear operation where much higher efficiencies can be obtained.

One or two of the simpler causes and results of nonlinear operation can be briefly indicated by means of Figs. 7.9 and 7.10. Neglecting for the moment any resistance in the grid circuit r_c, Fig. 7.9 shows the curvature of the effective plate characteristics.

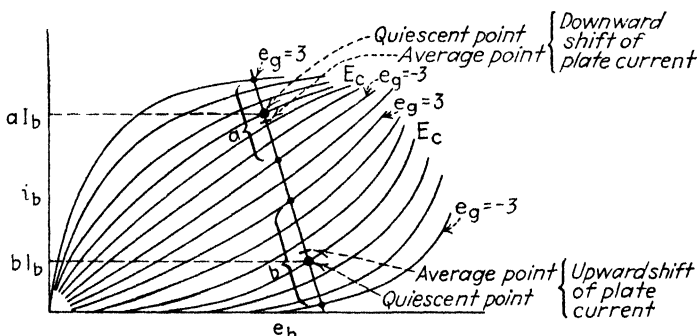



FIG. 7.9.—Distortion in nonlinear regions.

Consider the operating region a determined by, say, a sinusoidal $e_0 = e_g$ with the maximum value $E_{gm} = 3$. On sketching out the plate-current wave it is seen that the average value lies below the quiescent value I_b, the wave is no longer sinusoidal, the upper loops being considerably flattened. In the case of operation in region b the reverse is seen to be the case. The lower loops of the plate-current wave are flattened and the average value is greater than the quiescent value I_b. In the limit of the quiescent point being practically down on the e_b axis, the lower loops are completely suppressed and the upper ones greatly distorted. It may be seen qualitatively that

very high efficiencies become possible. If any grid current is drawn and r_c is not negligible, the variation e_g will not be the same as e_0, the potential wave applied to the grid circuit. The maximum positive-grid excursion will be smaller than the maximum positive value of e_0 by the drop in the grid resistance $I_{gm}r_c$. This situation is illustrated in Fig. 7.10. The upper loops of i_p will be flattened and the average value will be shifted below I_b. These various phenomena are all frequently met in practice.

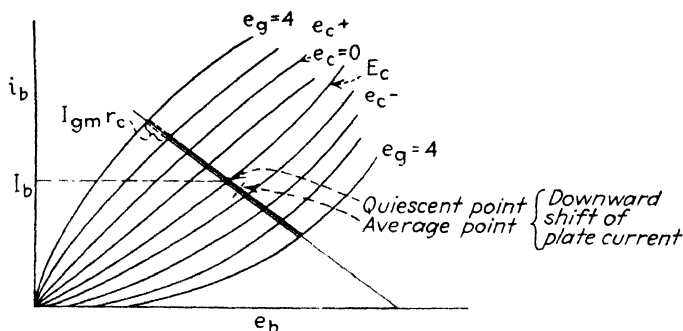



FIG. 7.10. Shortening of the operating path produced by the flow of grid current.

7.3. Multielectrode Tubes.—The triode is the prototype of all high-vacuum control tubes. These tubes are adaptable to many purposes some of which will be discussed in later sections of this chapter and in succeeding chapters. They are built for a wide range of powers from the small radio-receiving tube designed for a few milliwatts to the large tubes with water-cooled plates and grids which handle many kilowatts. It is in general only the small tubes that are designed for linear operation. The characteristics depend largely on the position and spacing of the grid. A coarse-mesh grid near the plate has relatively little control of the plate current and yields a small amplification factor, of the order of 1–5, whereas a fine-mesh grid close to the cathode has a much greater control over the space current and leads to amplification factors of the order of 20–100. The principal disadvantage of a triode, particularly in the case of those with large amplification factors, is the reaction of the plate potential upon that of the grid. The conductive reaction is indicated by the primed equations of the preceding section, but this is only important when a grid current exists. In addition there is an interaction between these two elements owing to the electrostatic capacity between them. This is in general undesirable and can be greatly reduced by the insertion of a second grid between the control grid and plate, which acts as a screening electrode isolating these elements from one another. Such a grid is known as a *screen grid* and a tube containing these four elements is called a *tetrode*. In a tetrode the plate-grid reaction, which couples these two circuits together and may give rise to undesired oscillations, is reduced to a negligible amount.

In ordinary operation the screen is kept at a fixed positive potential of the order of that of the plate E_b. This rather large positive potential is necessary since the cathode as well as the control grid is screened from the plate, and it is the screen potential that overcomes the space charge and allows a current to flow to the screen and plate. This positive screen potential, however, brings with it certain disadvantages. The electrons from the cathode strike the plate with an energy ee_b. Most of this energy is dissipated in heating the anode, but some of it may be transferred to conduction electrons near the surface of the plate and produce secondary emission (Sec. 6.6). If there is a retarding field

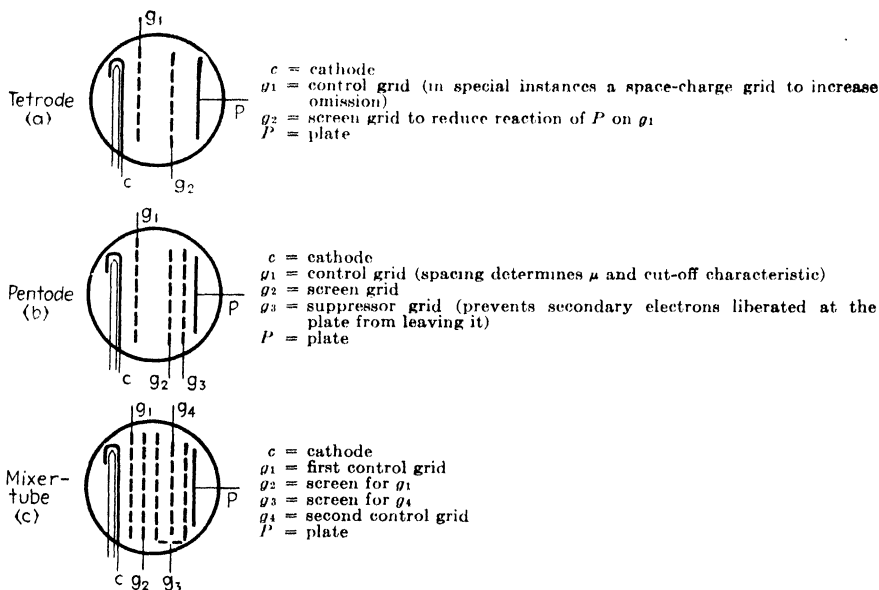



FIG. 7.11.—Typical multielectrode tubes.

at the surface of the plate, these secondary electrons return to it and give rise to no current in the external circuit. If, however, owing to the plate potential fluctuation, e_p, e_b is less than the screen potential, these electrons are drawn from the plate to the screen and this decreases the total electron current to the plate. The number of secondaries emitted per primary incident on the plate is a function of e_b and the nature of the plate surface; the fraction escaping from it to the screen is a function of the potential difference between these two elements and the potential distribution between them. In a typical tube these relations are such as to give rise to a net plate current of the form of curve 5 in Fig. 7.12. This characteristic exhibits a region with a negative slope which from Sec. 5.6 indicates instability unless the external circuit

contains a larger positive series resistance. This region is made use of in various circuits, but for general amplification work it is a disadvantage in limiting the useful linear range of the $i_b(e_b, e_c)$ surface.

The current due to secondary emission can be suppressed by the insertion of a third grid between the screen and plate which is maintained at a negative potential with respect to the latter and hence returns the secondary electrons to the plate from which they are liberated. This grid is known as a *suppressor grid* and the tube containing it is called a *pentode*. It is not necessary actually to insert a third grid to suppress the secondary current. With a suitable design of the tetrode elements an electron space charge can be built up between the screen and plate which alters the potential distribution in this region in such a way as to prevent electrons leaving the plate. This type of tube is known as a *beam tube* in view of the well-defined electron beams which pass through the open spaces of the grids from cathode to plate.

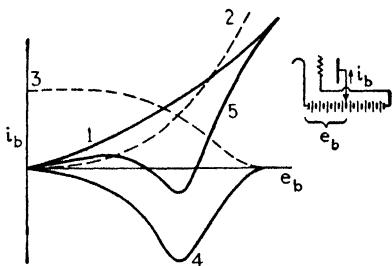



FIG. 7.12.—Effect of secondary emission on plate current. 1 = electron current from cathode to plate; 2 = secondary electrons liberated at plate; 3 = fraction of these that leave plate; 4 = 2 × 3, electron current leaving plate; 5 = 1 + 4, net plate current.

Figure 7.13 indicates the forms of typical characteristics in these different types of tubes. The triode curve resembles the three halves' power law in being concave upward, the others resemble more saturation curves and reflect the fact that the plate potential has little influence on the total space current. The dynamic plate resistance e_p/i_p is much larger over most of the diagram for these latter curves which is a disadvantage in that it requires the use of very high resistance loads for efficient power transfer or voltage amplification. The beam-tube characteristic exhibits the longest comparatively straight portion, its i_b, e_b diagram has the largest linear region, and hence it can handle linearly the largest grid potential fluctuations. Figure 7.14 indicates the different types of plate-current control exercised by the two standard types of pentode control-grid structures. This is an i_b, e_c diagram; it is a plane perpendicular to the e_b axis and the curves are the intersection of this plane, for some particular value of e_b, with the $i_b(e_c, e_b)$ surface. A uniformly spaced grid with equal control over the emission from all points of the cathode

FIG. 7.13.—Representative plate-circuit characteristics.

surface leads to the sharp cutoff type of control characteristic. The plate current becomes appreciable at a critical grid potential and rises approximately linearly with this variable over its useful range. It resembles the ideal rectifier characteristic (Fig. 5.16). The slope of this curve is s_p, and since it and r_p are constant over a large region of the $i_b(e_c, e_b)$ surface for this type of pentode, the amplification factor μ_p is also constant. If the grid spacing varies over different regions of the cathode surface, the degree of control exercised by it over the plate current will also vary and there will be no well-defined amplification constant μ_p. Very large negative values of e_c must be used to reduce i_b to a negligible amount, and s_p and μ_p vary widely over the useful range of e_c. Hence this type of tube is characterized by a *remote cutoff* or *variable* μ. It is very useful for many special nonlinear services.

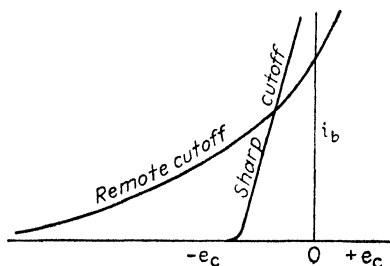



FIG. 7.14.—Typical pentode grid-potential–plate-current characteristics.

A large number of multielectrode tubes are merely combinations of diodes, triodes, etc., in the same envelope and as such contribute no new principles of operation. Also there are special-purpose tubes such as those employing positive or *space-charge grids* next to the cathode to increase the space current, but these are not of sufficient general interest to warrant further discussion. Furthermore, the types of tubes which have been discussed are not always used in the way which has been described. For instance, the nonlinear characteristic of the suppressor grid is frequently employed for modulation purposes. A refinement which is very useful in many instances for this purpose is the *mixer tube* of Fig. 7.11C. This has four grids, two of which are screens or shields and two are designed for control purposes. The control grids, which have different types of characteristics (Fig. 7.14), are shielded from one another and the plate. When potential waves are applied to these two grids, the plate current contains the sum and difference frequencies as well as harmonics, *i.e.*, the two frequencies are mixed.

7.4. Examples of the Use of Thermionic Tubes.—The principal uses of thermionic tubes can be roughly divided into three categories: amplification, modulation, and oscillation. The discussion of oscillation or the spontaneous conversion of direct-current energy supplied by the plate battery into alternating-current energy must be postponed till a later chapter. But as far as the tube is concerned, the necessary criterion for the performance of this function is the instability represented by a negative dynamic input or output resistance, *i.e.*, a falling characteristic as, for example, curve 5 of Fig. 7.12. This can be brought

about in various ways and the resultant phenomena will be discussed later. Modulation also involves the nonlinear character of the vacuum tube. The general principles of the process have already been described, and for the details of practical methods reference should be made to treatises on radio technique. The uses of tubes for amplification are divided into various classes. When the tube is used only in its linear region, the type of amplification is called *class A*. The voltage amplification obtained is given by Eq. (7.11) and the load current or potential drop reproduces exactly, though to a different scale, the potential applied to the grid. As has been previously mentioned, such a device is extremely inefficient as a power amplifier. If E_c is sufficiently negative (so-called *negative bias*) so that a negligible plate current flows in the absence of grid excitation ($e_g = 0$), the amplifier is known as *class B*. As has been pointed out in connection with Fig. 7.9, this leads to a nonlinear relation between i_p and e_g. However, the distortion so introduced can be largely eliminated by the use of tubes with suitable characteristics in the balanced circuit which will be described later. Reference to Fig. 7.9 indicates that this method of use leads to a much higher efficiency for power amplification. Finally, if the grid is biased very negatively so that plate current flows only for large positive values of e_g, the amplifier is known as *class C*. The discussion of this type will be postponed till a later chapter as its chief use is with a tuned load as a power amplifier at radio frequencies. In this application it has a very high efficiency. Multielectrode tubes not only amplify but also isolate the plate and grid circuits, avoiding the many undesired effects of interaction; these are frequently known as *buffer amplifiers*. Finally, a series of stages of amplification can be used in cascade to enhance the effect of a single tube. One or two of the methods of connecting or coupling successive plate and grid circuits will be mentioned later.

The thermionic amplifier resembles the electrometer in that it is characterized by a high input resistance, *i.e.*, e_g/i_g is large. For this reason it is suitable for the measurement of very small currents. The grid current is, of course, never strictly zero. At high frequencies the grid capacity is important, but even in low-frequency and static uses the grid conduction current is a limiting factor. There is some residual gas which gives rise to a positive-ion current to the grid; these ions may liberate secondary electrons which are drawn away and contribute to the grid current. Also the light and heat from the cathode may give rise to photoelectric and thermionic electron emission which further enhances the grid current. In the average tube this current is very roughly of the order of 10^{-10} amp. and is subject to fluctuations of a somewhat smaller order. This places an upper limit of about 10^8 ohms to the resistance that can be employed in the grid circuit. If special precautions

are taken in the construction and use of the tube, the grid current can be made very small and its fluctuations reduced to about 10^{-15} amp.; this means that currents of this order of magnitude can be measured and tubes of this type, such as the FP-54 and the Victoreen VX-41, are used for electrometer work. Grid-current fluctuations impose a limit on all small signal amplification, for they are amplified in succeeding stages, as is the desired signal, and give rise to background noise.

The high resistance associated with a vacuum tube is a disadvantage in that it limits its efficient use to high-resistance circuits. Consider an emf. e_0 generated in a resistance R_0. If a galvanometer is placed in series as in a of Fig. 7.15, the deflection is given by (Sec. 10.5)

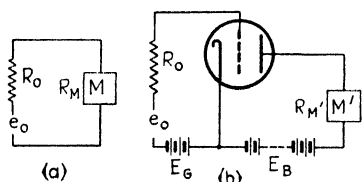



FIG. 7.15.—Schematic circuits illustrating the use of a vacuum tube as an amplifier.

$$D = c\sqrt{R_M}i = c\frac{\sqrt{R_M}}{R_0 + R_M}e_0$$

where c is a constant of the galvanometer. For the greatest efficiency

$$R_0 = R_M \qquad \text{and} \qquad D = c\frac{1}{2\sqrt{R_0}}e_0$$

If a tube is employed as shown in b of Fig. 7.15

$$D' = c'\sqrt{R_{M'}}i_b$$

Assuming a circuit arrangement in which the constant plate current I_b does not flow through the meter

$$D' = c'\sqrt{R_{M'}}i_p = c'\frac{\mu_p\sqrt{R_{M'}}}{r_p + R_M}e_0$$

For the optimum case of $r_p = R_{M'}$

$$D' = c'\frac{\mu_p}{2\sqrt{r_p}}e_0$$

Hence, if c and c' are the same

$$\frac{D'}{D} = \mu_p\sqrt{\frac{R_0}{r_p}}$$

Even assuming the most suitable meters are available, a tube can only be used to advantage if the quantity on the right is greater than unity. With, for example, a μ_p of 25, R_0 cannot be less than $r_p/625$ if there is to be any advantage in the use of a tube. As r_p is, say, of the order of 10,000 ohms, this implies that R_0 must be greater than about 15 ohms. Thus voltaic-cell and thermocouple potentials cannot in general be

amplified directly in this way. On the other hand, photoelectric currents are ideally suited for tube amplification as they are developed in high-resistance circuits. Figure 7.16 indicates a typical simple circuit. The iR drops in the resistances 1-4 (known as a *bleeder resistance* or *voltage divider*) provide the tube and photocell potentials. If the cell is placed in position a, illumination decreases the plate current; if it is in position b, the plate current is increased by illumination.

The high input resistance makes the vacuum tube useful as a voltmeter, for its grid circuit draws very little current and hence alters but slightly the circuit under measurement. A discussion of a typical voltmeter circuit, such as that of Fig. 7.17, illustrates many points of interest in vacuum-tube technique. In the first place, all the elements,

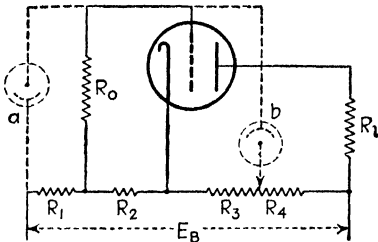



Fig. 7.16.—Photoelectric cell and amplifier circuit.

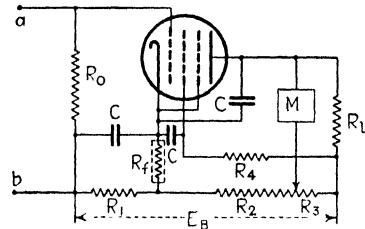



Fig. 7.17.—Vacuum-tube-voltmeter type of circuit.

with the exception of the control grid, are connected to the cathode through condensers. These present a low-resistance path to rapid fluctuations of the potential of these elements; hence the tube is effectively short-circuited except for relatively slow variations. A similar by-pass condenser connects the lower end of the grid resistance to the cathode. The static plate current I_b does not flow through the meter M because what is essentially a bridge circuit is employed. The lower meter terminal is adjusted along the bleeder resistance until the following relation is fulfilled:

$$\frac{\text{Tube drop } (E_b)}{\text{Load drop } (I_b R_l)} = \frac{\text{drop in } R_2}{\text{drop in } R_3} \qquad i.e., \qquad \frac{R_p}{R_l} = \frac{R_2}{R_3}$$

where $R_p = E_b/I_b$, the effective static plate resistance. (R_f is for the moment neglected.) In this condition there is no potential drop across the meter and hence no current through it. If the grid potential is altered the balance condition is disturbed and the meter deflects. From Sec. 4.6 it is evident that R_l and the meter resistance should be of the order of the plate resistance for maximum sensitivity, but this condition is seldom realized in practice. It might be mentioned that for the balance condition to be preserved for fluctuations in the supply voltage E_B, R_p should equal r_p. This condition seldom holds but the inclusion

of additional resistances can effectively bring it about. This dynamic balance is automatically achieved in the push-pull type of circuit to be described later.

The type of control exercised by the grid determines the action of the tube. Assume first that it is the sharp cutoff type biassed to the cutoff point. Positive-grid excursions are linearly reflected in the plate current and negative ones are ineffective which for a sinusoidal e_g leads to an i_p resembling Fig. 5.17. Hence the meter deflection which is proportional to the direct-current component of i_p is also proportional to the average of the positive values of e_g. Thus this arrangement measures the average positive-grid potential. If the meter terminals are interchanged, the average value of the loops that were previously the negative ones are measured. Hence, if the wave is unsymmetrical, a different meter reading will result on reversal and some information about the wave shape can be obtained. The sensitivity can be reduced by the inclusion of a cathode resistance R_f[1]. In this case it can be shown that the e.p.c. theorem becomes

$$i_p = \frac{\mu_p e_0}{(l + \mu_p)R_f + r_p + r_l}$$

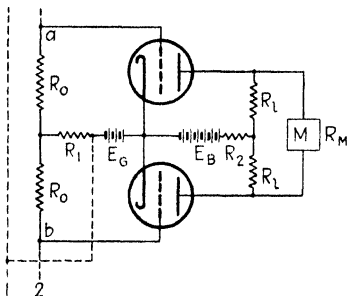



FIG. 7.18.—Push-pull amplifier and wattmeter type of circuit (modification for wattmeter indicated in dashed lines).

This value of i_p is smaller than that of Eq. (7.10) hence the meter deflection is less for the same average value of e_0, the potential fluctuation across R_0. If the tube used in the voltmeter circuit is of the remote cutoff type and biassed to the center of an approximately parabolic region of its grid characteristic, it is not the average but the r.m.s. value of the grid potential that is measured. (R_f is, of course, absent.)

There is one further type of circuit which deserves to be mentioned because of its high degree of symmetry and general usefulness. This is the *push-pull* circuit of Fig. 7.18. Its name is derived from the fact that a potential difference established between *a* and *b* has an opposite effect on the two grids, one plate current increases and the other decreases. The circuit may be used linearly as a voltmeter or balanced amplifier. If the tubes are identical, it can be seen by symmetry that the e_b's are the same for $e_g = 0$ and the meter is across a balanced bridge, two arms of which are the tubes and the other two the load resistances R_l. The

[1] In amplifier circuits the grid bias is frequently obtained by such a cathode resistor (bias = I_bR_f). In order that it shall not reduce the sensitivity in that case it is by-passed by a condenser which presents a low-resistance path to the frequency to be amplified.

bridge is also dynamically balanced against fluctuations in the plate and grid batteries. There are a number of other applications of this circuit when the tubes are operated beyond their linear range. Consider for simplicity that the tubes are pentodes so that the effect of e_p is negligible, the characteristic of each tube is then:[1]

$$i_p = s_p e_g + \frac{1}{2}\frac{d^2 i_b}{de_c^2}e_g^2 + \frac{1}{6}\frac{d^3 i_b}{de_c^3}e_g^3 + \cdots$$

If $2e_0$ represents the potential fluctuation between a and b, $e_{g1} = +e_0$ and $e_{g2} = -e_0$. Inserting these values in the characteristic, the potential fluctuation across the load is given by

$$\begin{aligned} e_l &= R_l(i_{p1} - i_{p2}) \\ &= 2R_l\left(s_p e_0 + \frac{1}{6}\frac{d^3 i_b}{de_c^3}e_0^3 \cdots\right) \end{aligned}$$

Thus the terms that give rise to even harmonics are absent. This is characteristic of the push-pull circuit and enables suitable tubes to be used in class B service with negligible harmonic distortion. If a variable emf. e_1 is developed across R_1 as well as e_0 across R_0, the circuit represents a balanced modulator. In this case

$$e_{g1} = e_1 + e_0 \qquad \text{and} \qquad e_{g2} = e_1 - e_0$$

The potential fluctuation across the load is

$$e_l = 2R_l\left(s_p e_0 + \frac{d^2 i_b}{de_c^2}e_0 e_1 \cdots\right)$$

The potential fluctuation across R_2 is $e_2 = R_2(i_{p1} + i_{p2})$ or

$$e_2 = 2R_2\left(s_p e_1 + \frac{1}{2}\frac{d^2 i_b}{de_c^2}(e_0^2 + e_1^2) \cdots\right)$$

If, for example, e_1 is a radio frequency and e_0 an audio one, e_l contains only the audio frequency and the side bands, the carrier e_1 having been eliminated. The carrier and even harmonics appear across R_2. The elimination of undesired frequencies represents a great saving of power in subsequent stages of amplification. The dashed lines in Fig. 7.18 indicate a low-power wattmeter circuit, the analysis of which is reserved for a problem.

7.5. Measurement of Tube Coefficients.—The method of obtaining the static curves of Fig. 7.6, for instance, is obvious. A milliammeter

[1] This expression is also applicable to a triode if it is considered as a composite characteristic, the differential coefficients being determined by both the grid and plate circuits.

in the plate circuit measures the plate current i_b as a function of the plate potential e_b for a series of values of the grid potential e_c. In obtaining these curves care must be taken not to exceed the power rating of either the plate or grid. These static characteristics are of great importance in determining the behavior of a tube, but they do not give all the information necessary for the proper design of tube circuits for alternating-current work. In the first place, as there is generally a high resistance in the grid circuit, a knowledge of the behavior of the grid current is important. Also, for high-frequency work the interelectrode capacities play an important role. For a discussion of the methods of measuring these parameters reference should be made to treatises on this subject.

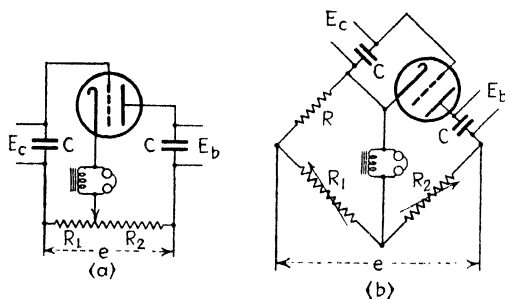



FIG. 7.19.—(a) Simple circuit for measuring μ. (b) Simple circuit for measuring r_p.

The linear tube parameters r_p, s_p, and μ_p can be obtained approximately from the static characteristics or μ_p and r_p (and hence s_p through the relation $\mu_p = r_p s_p$) can be measured directly by means of the circuits shown in Fig. 7.19. The quiescent values of E_c and E_b are applied across the large condensers C which constitute a low-resistance path for the alternating current produced by the sinusoidal emf. e. The inductance across the phones is a low-resistance path for the direct tube currents and a high resistance for the varying component. In the circuit for measuring μ, e_g is proportional to R_1 and e_p to R_2; therefore, for no signal in the phones ($i = i_p = 0$) μ_p is given by $\mu_p = (e_p/e_g)_{i_p=\text{const.}} = R_2/R_1$. The circuit for measuring r_p is a bridge circuit in which r_p forms one arm. In balance $r_p = RR_2/R_1$. Obviously for optimum sensitivity all the resistances in both circuits (including that of the phones) should be of the same order of magnitude. For methods of measurement, taking into account the tube capacities, reference should be made to special treatises.

7.6. General Alternating-current Theory of Resistance-capacity Circuits.—For the analysis of the simplest and one of the most useful methods of cascade coupling as well as for the clarification of some of the statements made in recent sections it is necessary to discuss the use of condensers in alternating-current circuits. As the general theory

can be handled adequately on the basis of concepts that have been previously discussed, it will be introduced at this point. Figure 7.20 represents diagrammatically a resistance R and a capacity C in series with an alternating potential V. Applying Kirchhoff's law

$$V = V_C + V_R = \frac{q}{C} + iR$$

Writing $i = dq/dt$, this equation can be written in either of the following forms:

$$R\frac{dq}{dt} + \frac{q}{C} = V \tag{7.12}$$

or

$$R\frac{di}{dt} + \frac{i}{C} = \frac{dV}{dt} \tag{7.12'}$$

Both of them are of the form of Eq. (C.5) (Appendix C) and the solution is given by Eq. (C.6). As only the steady-state solution will be discussed here (the transient term containing α is neglected), a very simple and convenient solution will be obtained by an alternative method employing the representation of i and V as complex quantities. If V is a sinusoidal function of ωt, it may be considered to be either the real or imaginary part of the complex emf. $\mathbf{V}$, where

FIG. 7.20.—Alternating-current circuit containing both resistance and capacity.

$$\mathbf{V} = V_0(\cos \omega t + j \sin \omega t) = V_0 e^{j\omega t}$$

(see Appendix C). $\mathbf{V}$ itself is, of course, not an acceptable emf., for it is complex, and an applied emf. is a real quantity, but the introduction of $\mathbf{V}$ is a great convenience in handling problems and the actual values of V and i are obtained by accepting only the real part or imaginary part, whichever has been previously designated. Substituting $\mathbf{V}$ for V in Eq. (7.12′) and assuming the solution $\mathbf{i} = i_0 e^{j\omega t}$, the equation becomes

$$i_0\left(Rj\omega + \frac{1}{C}\right)e^{j\omega t} = j\omega V_0 e^{j\omega t}$$

Hence the complex representation of the current $\mathbf{i}$ is given by

$$\mathbf{i} = \frac{\mathbf{V}}{\left(R + \frac{1}{j\omega C}\right)} = \frac{\mathbf{V}}{\mathbf{z}} \tag{7.13}$$

where $\mathbf{z}$ is known as the complex *impedance*. This linear relation between $\mathbf{i}$ and $\mathbf{V}$ which is due to the linearity of the differential equation [Eq. (7.12′)] has very important consequences. The relation between $\mathbf{i}$

and **V** is the same as the ohmic one between i and V with **z** replacing R. All the general resistance theorems which were developed in Sec. 4.3 evidently hold equally well for **i** and **V** in resistance-capacity circuits if the complex impedance **z** of each circuit branch is used in place of R. Thus the formal treatment of these circuits is identical with that of pure resistance circuits.

Equation (7.13) is frequently written in different ways by means of the complex identities of Eq. (C.14). Thus, if X, the *capacitative reactance*, is written for $-\frac{1}{\omega C}$, $\mathbf{z} = R + jX = Ze^{j\varphi}$, where $Z = (R^2 + X^2)^{1/2}$

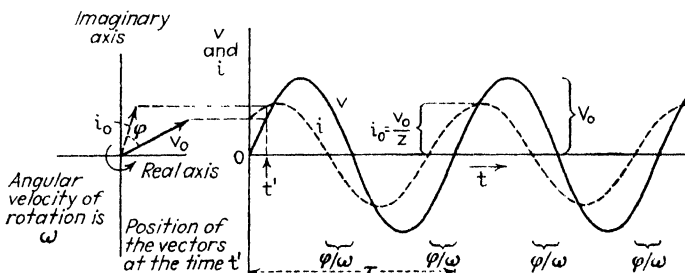



FIG. 7.21.—Potential and current in a resistance-capacity circuit.

and $\varphi = \tan^{-1} \frac{X}{R}$. Z is the absolute magnitude of the impedance. Then

$$\mathbf{i} = \frac{V_0}{Z} e^{j(\omega t - \varphi)}$$

Or if the real part of **V** and **i** are taken as representing the actual value of the potential V and current i

$$i = \frac{V_0}{Z} \cos (\omega t - \varphi) \tag{7.14}$$

A sinusoidal potential and current wave typical of this type of circuit are shown in Fig. 7.21. The magnitude of the current wave is less than that of the potential wave by the factor $1/Z$; also the current wave is shifted along the axis of increasing time by an amount φ/ω, or, since X is negative, φ is negative and the shift is to the left, *i.e.*, to an earlier time. In the case of a resistance-capacity circuit the current is said to *lead* the potential by a time φ/ω, or by the phase angle $\varphi = \tan^{-1} (1/R\omega C)$. The left-hand portion of Fig. 7.21 indicates the complex representation of the potential and current. V_0 and i_0 are the magnitudes of the two vectors in the complex plane. They make an angle φ with one another and because of the common term $e^{j\omega t}$ they may be considered to rotate about the origin with an angular velocity ω in a counterclockwise sense

(multiplication by $e^{j\theta}$ represents counterclockwise rotation through the angle θ), retaining their relative positions. If the imaginary parts of **V** and **i** represent the actual potential and current, V and i are given by the projection of **V** and **i** on the imaginary axis and the curve at the right in the figure results. In the case of a pure resistance the capacity is effectively infinite and $\varphi = \tan \varphi = 0$, *i.e.*, **V** and **i** are colinear and the waves are in phase. In the case of a pure capacity the resistance is zero and $\tan \varphi = \infty$ or $\varphi = \pi/2$, *i.e.*, **V** and **i** are at right angles (**i** being along the positive j axis when **V** is along the positive real axis) and the current wave leads the potential wave by 90°.

If V is sinusoidal, i is also sinusoidal, *i.e.*, there is no distortion of a sinusoidal wave. The ratio of the maximum and effective values is the same. $V_0/i_0 = V_e/i_e = Z$. But Z is a function of the frequency. For a pure capacity $Z = 1/\omega C$, and if ωC is large, the current is large or the potential drop across the condenser is small. Thus, for sufficiently great values of the product ωC the capacities in Figs. 7.17 and 7.19 have a negligible alternating-current drop across them. If the potential wave applied to an R-C circuit is not sinusoidal, the current wave will not resemble that of the potential but will be distorted. The type of distortion can be seen immediately from a Fourier analysis of V. A term in the series of the form $V_n \cos n\omega t$ will give rise to a term in the series representing the current, which leads this potential term by the angle φ_n, and i_n, the magnitude of this current term, is given by V_n/Z_n, where $\varphi_n = \tan^{-1} \dfrac{1}{n\omega RC}$ and $Z_n = \left(R^2 + \dfrac{1}{(n\omega C)^2}\right)^{1/2}$. Thus successive terms having different n's will be affected differently and distortion will result.

It has previously been seen that the average consumption of alternating-current power is given by $i_e V_e \cos \varphi$; hence, as the phase difference between the current through a pure condenser and the potential across it is $\pi/2$, the power dissipated in it is zero. The situation can be analyzed in more detail if Eq. (7.12) is multiplied by $i = dq/dt$.

$$R\left(\frac{dq}{dt}\right)^2 + \frac{1}{C} q \frac{dq}{dt} = iV = P$$

The first term is always positive, whereas the second is as often positive as negative during a cycle; thus all the power is dissipated in R and none in C. The average power dissipation is given by integrating over a period and dividing by τ. Writing $V = V_0 \cos \omega t$ and $i = i_0 \cos (\omega t - \varphi)$

$$P_R = \frac{Ri_0^2}{\tau} \int_0^\tau \cos^2 (\omega t - \varphi) dt = R\frac{i_0^2}{2} = Ri^2$$

$$P_C = \frac{l}{\tau C} \int_0^\tau q\, dq = 0$$

Of course, all condensers are not perfect. If the dielectric has some conductivity or if owing to molecular friction all the energy of charge is not liberated on discharge, power is dissipated in the condenser.

The power loss that occurs in a condenser depends on the nature of the dielectric and the frequency (Sec. 3.2). If the dielectric has an appreciable conductivity, the condenser can be represented to a first approximation by a perfect condenser C and a parallel resistance R_P. The complex impedance of this circuit is $R_p/(1 + j\omega CR_P)$, and as R_P must be very large for any acceptable condenser, the power factor is approximately $1/\omega CR_P$. For a constant R_P this evidently decreases with increasing frequency and the characteristics of the condenser improve. A more important phenomenon in practical condensers is that of dielectric absorption. The characteristics of most dielectrics are such as to simulate to a first approximation a condenser in series with a comparatively small resistance. The charge on the condenser does not attain its maximum value the instant a potential is applied, but the charging current may be appreciable over a period of minutes. Likewise the charge does not flow off instantaneously when the terminals are shorted; there is a certain time lag associated with the dielectric which gives rise to a hysteresis phenomenon. If the instantaneous value of the charge is plotted for a complete period as a function of the potential, the curve is not closed and the area represents the energy dissipated in the dielectric in one cycle. The approximate equivalent circuit is that of Fig. 7.20, for which the power factor is approximately $R_s\omega C$, where R_s is the series resistance, if the power factor is small. However, the effective series resistance decreases with the frequency in such a way that over a very large range of this variable the power factor is approximately a constant characteristic of the dielectric material. This power loss, whether due to leakage or dielectric absorption, raises the temperature of the dielectric with consequent changes in its characteristics and generally a decrease in its dielectric strength. Thus condensers designed for direct current frequently fail on alternating current.

FIG. 7.22.—Condenser with a partially conducting dielectric.

Static methods of measuring capacities have been discussed in a much earlier section, but alternating-current methods are generally more convenient for routine measurements, provided the capacity is designed for alternating-current work. Large capacities can be measured in a power line with a series alternating-current ammeter and shunt alternating-current voltmeter just as a resistance of moderate value would be measured in a direct- or alternating-current circuit. A more

accurate method for the general measurement of capacities is supplied by a bridge circuit. A capacity bridge is indicated schematically in Fig. 7.23. In general two arms are pure resistances and the other two are pure capacities or capacities shunted by other resistances. An alternating potential is applied to the bridge and a suitable detecting device takes the place of the galvanometer in a direct-current bridge. Ordinary power frequencies are suitable for the measurement of large capacities and a vibration galvanometer, tuned to the power frequency, provides a sensitive detecting device. For smaller capacities audio and even radio frequencies are employed and an amplifier and phones or meter are used to indicate the balance point. The analysis of the bridge circuit in balance is very simple. The condition for a zero potential difference across the output is

FIG. 7.23.—Capacity bridge.

$$\frac{\mathbf{z}_1\mathbf{i}_1}{\mathbf{z}_2\mathbf{i}_2} = \frac{\mathbf{z}_3\mathbf{i}_3}{\mathbf{z}_4\mathbf{i}_4}$$

Or, since $\mathbf{i}_1 = \mathbf{i}_2$ and $\mathbf{i}_3 = \mathbf{i}_4$, the complex impedances must be in the ratio $\mathbf{z}_1/\mathbf{z}_2 = \mathbf{z}_3/\mathbf{z}_4$. As the equality of complex quantities implies the equality of both real and imaginary parts, the following equations result on using the constants of Fig. 7.23.

$$\frac{R_1}{R_2} = \frac{R_3}{R_4} = \frac{C_2}{C_1}$$

The first may be considered as the direct-current and the second the alternating-current condition. To obtain both conditions either the numerator or denominator in two of the fractions must be variable. For instance, C_2 may be variable instead of R_1, as shown in Fig. 7.23. The sensitivity conditions are the same as those of Sec. 4.6 with the substitution of Z's for the R's.

7.7. Resistance-capacity-coupled Alternating-current Amplifier.—For many purposes one stage of vacuum-tube amplification is insufficient and a series of units must be used in cascade. For direct-current amplification this implies a direct connection between the plate of one tube and the grid of the succeeding one and also a high-potential supply for all the tubes which are effectively in series. This arrangement has numerous disadvantages and it is unnecessary for alternating-current work. In this service any of a number of reactive coupling methods can be used. These often have frequency-response advantages and also enable the isolation of individual stages. Separate low-potential supplies may be used for each stage, or by a suitable decoupling circuit the

same supply may be used for a number of stages. Such a decoupling net is indicated schematically in the lower part of Fig. 7.25. The one shown employs resistances and capacities, but in practice choke coils (high alternating-current resistance, low direct-current resistance) replace the resistances, R, in current-carrying branches. The capacities C are of such a value as to present a low impedance to the alternating current which is to be amplified. Hence the lower ends of R_0 and R' and the screen grid are all effectively short-circuited to the cathode for this frequency. The resistances R isolate the upper circuit from the bleeder potential supply. Roughly speaking, the upper C's reduce the reaction of the tube on the potential supply and the lower ones reduce the reaction of any alternating current in this supply on the tube. Thus a number of tubes may receive their potential supplies from the same E_B if a decoupling circuit of this type is employed.

FIG. 7.24.—Resistance-capacity-coupled cascade amplifier unit.

One of the most satisfactory methods for the cascade connection of stages in a low-power linear amplifier is so-called resistance-capacity coupling. A unit of such an amplifier is shown in Fig. 7.24. The connection between the plate of one stage and the grid of the next is made through the capacity C. Thus fluctuations in e_{p1} are transmitted to e_{g2}, and so on—down the chain. The analysis of the operation of such an amplifier provides an instructive exercise in tube- and resistance-capacity circuits and yields useful information as to its frequency response characteristic. For this purpose the interelectrode tube capacities must be explicitly recognized. Consider that the load is made up of the plate resistance R_l, the coupling capacity C, and R_0 and that C_g is the effective capacity of the subsequent stage. Then by Kirchhoff's law

FIG. 7.25.—Schematic decoupling net to enable a number of tubes to obtain their power from one source.

$$\frac{\mathbf{e}_l}{\mathbf{z}_l} = \mathbf{i}_l = \mathbf{i}_p + \mathbf{i}_{fp} - \mathbf{i}_{pg}$$

and as

$$\mathbf{i}_p = s_p\mathbf{e}_g + k_p\mathbf{e}_p$$
$$\mathbf{i}_{fp} = j\omega C_{fp}\mathbf{e}_p \qquad \text{and} \qquad \mathbf{e}_p = -\mathbf{e}_l$$
$$\mathbf{i}_{pg} = j\omega C_{pg}(\mathbf{e}_g - \mathbf{e}_p)$$

substituting these values

$$\frac{\mathbf{e}_l}{\mathbf{e}_g} = \frac{(s_p - j\omega C_{pg})}{(\mathbf{y}_l + k_p + j\omega(C_{fp} + C_{pg}))}$$

Here $\mathbf{y}_l$, which is known as the *admittance*, is written for $1/\mathbf{z}_l$. Analyzing the load, the drop across $a'b'$, which is written $\mathbf{e}_g'$, is to the drop $\mathbf{e}_l$ across R_l as:

$$\frac{\mathbf{e}_g'}{\mathbf{e}_l} = \frac{j\omega C}{k_0 + j\omega(C + C_g)}$$

where the conductance k_0 is written for $1/R_0$ and C_g is the effective capacity presented by the grid-cathode terminals.[1] The product of these two expressions is $\mathbf{e}_g'/\mathbf{e}_g$ the voltage amplification per stage. The voltage amplification for n stages is this quantity raised to the nth power.

The voltage amplification per stage (V.A.) calculated from the above expressions for the general case is very involved. But in practice a considerable simplification is possible. Consider, for example, the pentode 6-J-7. For this tube $C_{pg} = 5 \times 10^{-15}$ farad and $s_p = 10^{-3}$ mho; therefore below frequencies of the order of 10^{10} the second term in the numerator of the first expression can be neglected in comparison with the first. Also the coupling capacity C is generally large in comparison with the interelectrode capacities, which leads to further simplifications. At fairly large frequencies (where the first expression is the important one) $\mathbf{y}_l = k_0 + k_1 + j\omega C_g$, and in the denominator of the second expression C_g can be neglected in comparison with C. A numerical value for V.A. can be obtained if the other circuit constants are assumed. For the 6-J-7, $C_{fp} = 12 \times 10^{-12}$ and $C_{gf} = 7 \times 10^{-12}$, $\mu_p = 1{,}500$ and $r_p = 1.5 \times 10^6$; assume $R_1 = 10^5$, $R_0 = 10^6$, and $C = 10^{-7}$ ($\frac{1}{10}$ μf.). Multiplying the expressions together and calculating the absolute value

$$\text{V.A.} = \frac{1{,}500}{\left[(18)^2 + \left(3 \times 10^{-5}\omega - \frac{180}{\omega}\right)^2\right]^{1/2}}$$

This may profitably be plotted on a semilogarithmic scale. The maximum occurs for $\omega = 2{,}450$, or a frequency of about 390 cycles, where it has the value of 83. It is greater than $1/\sqrt{2}$ of this value for all frequencies from a few cycles up to about 10^5. The lower limit cannot be accurately calculated from this expression. In general linear amplifier

[1] For resistive loads and for values of the frequency very much less than the ratio of plate resistance to interelectrode capacities, $C_g = C_{fg} + C_{pg}\left(1 + \frac{\mu}{1 + r_p k_l}\right)$, where k_l is the load conductance. This may be derived from the equation for the input impedance in Sec. 15.6.

construction the same type of tube is not used throughout. For the amplification of very small signals the first tube should be chosen for its low random grid current. After this tube the maximum of undistorted amplification is the desired characteristic. The final tube is chosen to deliver the requisite amount of power to the output device. All stages should be well shielded from one another for any appreciable interaction may lead to spontaneous oscillations of the relaxation type. This phenomenon will be further discussed in connection with the general theory of unstable circuits.

Cathode-follower Amplifier.—It has been seen that the amplification of a resistance-coupled stage decreases in the high-frequency range. The input capacitative reactance of the following stage decreases and becomes much less than the output impedance of the first stage. The requirement of small input capacitance of this first stage dictates the use of a pentode

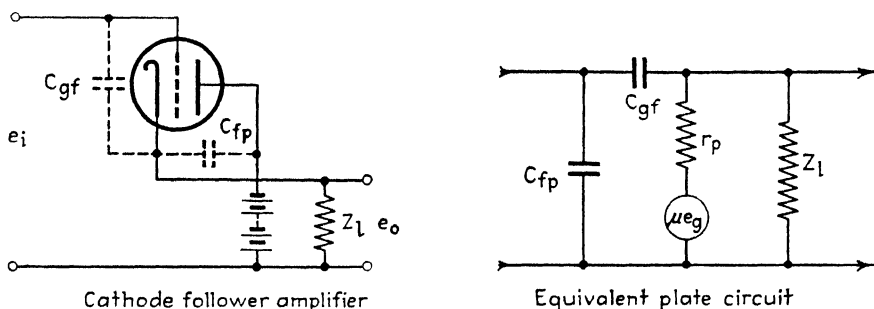



FIG. 7.26.—Cathode-follower amplifier.

whose output impedance is very high unless the coupling resistances are small. Small coupling resistances lead to low amplification and excessive distortion due to the curvature of the characteristic. This difficulty is removed by the use of two high-gain stages, usually pentodes, separated by a very low-gain triode stage in the form of a cathode follower, the basic circuit of which is shown in the figure.

Although the cathode follower has a voltage amplification less than unity, it has low input capacitance, high input impedance, low output impedance, and low distortion. It contributes nothing to the over-all amplification but makes possible higher uniform amplification over a given frequency range than can be obtained with conventional stages and plays an important role in wide-band high-gain linear amplifiers for physics research. It is often used as the last stage of an amplifier when the output voltage must be independent of the external impedance connected across the output of the amplifier. It is particularly useful in matching amplifier outputs to low-impedance devices over a wide frequency range. Since the output voltage is in phase with the input, it enjoys other special uses.

From solving the equivalent circuit of Fig. 7.26 the voltage amplification is given by

$$\text{V.A.} = \frac{(j\omega C_{gf}r_p + \mu)\mathbf{z}_l}{j\omega(C_{gf} + C_{fp})r_p\mathbf{z}_l + r_p + (\mu + 1)\mathbf{z}_l}$$

At frequencies where tube capacitances have little effect, this reduces to

$$\text{V.A.} = \frac{\mu\mathbf{z}_l}{r_p + (1 + \mu)\mathbf{z}_l}$$

which approaches the limit $\mu/(\mu + 1)$ as $\mathbf{z}_l/r_p$ becomes large. For ordinary tubes voltage amplification is generally 0.9 or more with an r_p of 10,000 ohms. The output terminal impedance is shown to be $r_p/(1 + \mu)$, being slightly less than the reciprocal of the transconductance. Transconductances are large enough so that it is generally only a few hundred ohms.

7.8. Electron Beams and Electrostatic Focusing.—There is a second very important type of thermionic tube which is used for the visual study of electrical phenomena. Its action depends on the fluorescent light that is emitted by many substances when they are bombarded by electrons with energies of several hundred volts. The glass envelope of discharge tubes and the older type of gas-filled X-ray tubes fluoresces with a greenish-yellow glow in those regions subject to cathode-ray (electron) bombardment. The nature of the light emitted depends on the material of the target. Electrons, which possess sufficient energy, drive some distance through the crystal lattice, disrupting the electron structure around the atomic lattice points. When this structure returns to its normal configuration, some of the excess energy is emitted in the form of radiation. Certain substances such as impure sulphides, borotungstates, etc., for reasons which are not well understood, fluoresce much more brilliantly than others. If these materials are coated on electrodes or on the inside of the glass envelope and electrons of several hundred volts energy directed upon them, they emit a brilliant blue or greenish-yellow light. Such coatings are known as fluorescent screens and because the intensity of the fluorescence from any small region is dependent on the number and energy of the electrons striking it, they provide a very valuable tool for the study of the motion of electrons in vacuum tubes. The material of the screen and the method of preparation determine the duration of the fluorescent light. This varies from a few microseconds to several seconds. The ones for which the fluorescence ceases almost as soon as the electron bombardment are most commonly used, as they are best adapted to the study of transient phenomena.

Fluorescent screens find many applications in electronic devices. The most important ones are in the class of cathode-ray oscillographs and television tubes which involve the production of a fine electron beam

or the focusing of a thermionic or photoelectric surface on the fluorescent screen. This type of control of electron motion can be effected most conveniently by means of a suitable arrangement of electrostatic fields. The actual design of focusing electrodes is largely empirical, but an analysis of the action of slits or diaphragms illustrates the general principles involved and explains the behavior of the simpler devices. Consider a flow of electrons from left to right through a diaphragm in the plane $x = 0$. For simplicity assume that it is a slit parallel to the z axis, which reduces the problem to a two-dimensional one in the xy plane. The electron density is generally small enough so that space charge can be neglected and Laplace's equation in two dimensions applied.

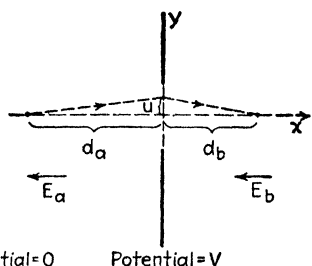



FIG. 7.27.—Change in an electron's trajectory on passing through a diaphragm from one electric field to another.

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0$$

The force on an electron in the y direction is given by $e\,\partial V/\partial y$ which can be expanded by Taylor's series about the origin

$$e\frac{\partial V}{\partial y} = e\left(\frac{\partial V}{\partial y}\right)_0 + ey\left(\frac{\partial^2 V}{\partial y^2}\right)_0 + \cdots$$

By symmetry there is no y component of force at the origin and the first term is zero. Considering the limited region near the origin for which $(\partial^2 V/\partial y^2)_0$ can be written for $\partial^2 V/\partial y^2$ and the square and higher power of y neglected

$$e\frac{\partial V}{\partial y} = -ey\left(\frac{\partial^2 V}{\partial x^2}\right)$$

This force is equal to the mass times the acceleration, $m\,d^2y/dt^2$. If the velocity, v, is considered constant in this small region

$$\frac{d^2y}{dt^2} = v^2\frac{d^2y}{dx^2}$$

(the path of the electrons is assumed to make but a small angle with the x axis so that v is approximately equal to its x component) and the equation becomes

$$\frac{d^2y}{dx^2} = -\frac{ey}{mv^2}\frac{\partial^2 V}{\partial x^2}$$

Assuming further that y is approximately constant and equal to u the

equation may be integrated to give

$$\left.\frac{dy}{dx}\right]_a^b = \left.\frac{-u}{2V}\frac{\partial V}{\partial x}\right]_a^b$$

where $eV = \frac{1}{2}mv^2$ is the energy possessed by the electrons when they reach the diaphragm which is at the potential V with respect to the emitting surface. Since $(dy/dx)_a = u/d_a$ (Fig. 7.27) and $(\partial V/\partial x)_a = -E_a$ where E is the electric field

$$\frac{1}{d_b} - \frac{1}{d_a} = \frac{E_b - E_a}{2V}$$

This is the same as the simple lens equation of geometrical optics with $2V/(E_b - E_a)$ for the focal length f. Thus to this degree of approximation the effect of a diaphragm on a beam of electrons is similar to the action of a thin lens on a beam of light. If the diaphragm had been considered to be circular, Laplace's equation in cylindrical coordinates would have led to a focal length of $4V/(E_b - E_a)$, or to cover these two cases

$$f = \frac{2gV}{E_b - E_a} \qquad (7.15)$$

FIG. 7.28.—Divergence produced in a parallel electron or ion beam on passing through a diaphragm into a field-free region.

where for slits $g = 1$ and for holes $g = 2$. Eq. (7.15) is seen to be independent of e and m and hence holds for any type of charged particles. If E_a is greater than E_b, f is negative and the lens is diverging. Thus the situation depicted in Fig. 7.28 of electrons drawn from a plane cathode through an opening into a field-free space always results in a divergent beam, the angle of divergence α being given by $u/2ga$. This is seen to be independent of V. However, if there is a field to the right of the diaphragm, a positive lens can be produced and electrons leaving the cathode normally can be focused on a fluorescent screen.

In a cathode-ray oscillograph and other similar devices it is desired to project a narrow parallel beam of electrons into a field-free region. This requires the use of two diaphragms. The arrangement is shown schematically in Fig. 7.29. The first diaphragm is at a potential V and the second at a potential MV. Between the two the electrons are accelerated at a constant rate along the x axis and hence move in a parabolic path. This parabola can be written

$$q + x = c(p - y)^2$$

The constants can be determined from the initial conditions at diaphragm

1 for which let $x = 0$. Here $y = u$, and for normal incidence the slope of the trajectory on leaving is $dy/dx = u/f_1$.

$$\frac{d^2x}{dy^2} = \frac{1}{v_y^2}\frac{d^2x}{dt^2} = \frac{1}{\frac{2eV}{m}\left(\frac{u}{f_1}\right)^2}\frac{eV(M-1)}{mb} = \frac{f_1^2}{2su^2}$$

where s is written for $b/(M-1)$. Determining the constants from these conditions the trajectory becomes

$$y = \frac{2su}{f_1}\left[\left(1 + \frac{f_1}{2s}\right) - \left(\frac{x}{s} + 1\right)^{\frac{1}{2}}\right] \tag{7.16}$$

At the second diaphragm, for which $x = b$, let y have the value w. For

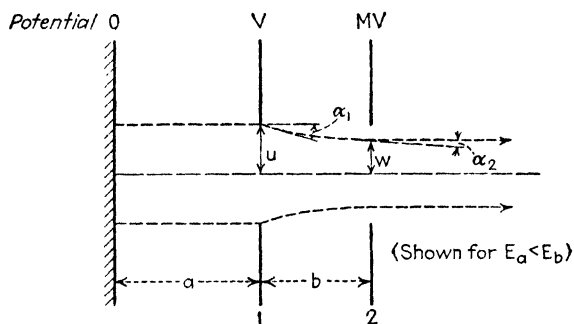



FIG. 7.29.—Production of a parallel beam of charged particles by a pair of slits or diaphragms.

collimation the angular divergence of the second diaphragm, w/f_2, must be equal to the entering angle of convergence $(dy/dx)_b$. Hence

$$\left(\frac{b}{s} + 1\right)^{\frac{1}{2}} - \frac{f_2}{2s}\left(\frac{b}{s} + 1\right)^{-\frac{1}{2}} = 1 + \frac{f_1}{2s}$$

Substituting:

$$f_1 = \frac{2gb}{(M-1) - \frac{b}{a}} \quad \text{and} \quad f_2 = -2gsM$$

The condition becomes

$$\frac{b}{a} = \frac{(M-1)(M^{\frac{1}{2}} - 1)}{M^{\frac{1}{2}} - \frac{1}{(g+1)}} \tag{7.17}$$

This is the relation that must exist between the ratio of the two separations and the relative potential of the second diaphragm for a parallel beam of electrons to be projected into the field-free space to the right of diaphragm 2. The width of this beam is w. In practice M should be

greater than 2. In an oscillograph the electrodes are generally cylinders rather than diaphragms. In this case the calculations are much more difficult and the adjustment for parallelism is made empirically.[1]

7.9. Cathode-ray Oscillograph.—One of the most important uses of the electron beams which were described in the previous section is in the cathode-ray oscillograph. The beam is projected between two sets of plates which are at right angles to one another before it impinges on the fluorescent screen. The arrangement is shown schematically in Fig. 7.30. If an electric field E_y exists between one set of plates which are perpendicular to the y axis, the electron beam in passing through them is accelerated in the y direction at a rate eE_y/m. If the plates extend a distance l and the end effects can be neglected, the emergent y component

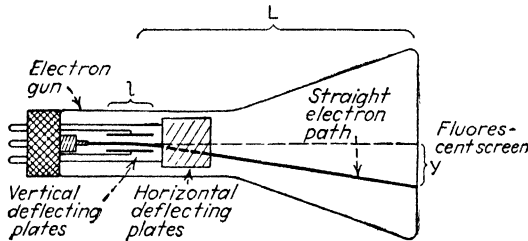



FIG. 7.30.—Cathode-ray oscillograph.

of the velocity is given by $eE_yt/m = eE_yl/mv$, where v, which is strictly the x component of the velocity, may be taken as equal to the total velocity if the energy gained in the field E_y is small in comparison with the energy eV of the electrons emerging from the diaphragm system. The tangent of the angle of emergence which is v_y/v_x is then eE_yl/mv^2, or writing $\frac{1}{2}mv^2 = eV$, the displacement of the intersection of the beam and the screen at a distance L from the plates is

$$y = \frac{Ll}{2V}E_y$$

As many simplifying assumptions have been made, this expression is only approximate, but it shows that the deflection of the spot on the screen is proportional to the field between the plates and inversely proportional to the energy of the beam. If the screen is equipped with a translucent scale, it can be calibrated and used as a voltmeter to measure the potential difference between the plates. The instrument is very useful in this capacity as a high-impedance voltmeter.

The other set of plates, of course, produces an independent deflection

[1] References: HANSEN and WEBSTER, *Rev. Sci. Instruments*, **7**, 17 (1936); KIRKPATRICK and BEKALEY, *Rev. Sci. Instruments*, **7**, 24 (1936); EPSTEIN, *Proc. I. R. E.*, **24**, 1095 (1936) ZWORYKIN, MORTON, RAMBERG, HILLIER, and VANCE," "Electron Optics and the Electron Microscope," John Wiley, Inc., New York, 1945.

of the beam along the z axis. If a translucent set of coordinates is placed over the screen, the two deflections can be compared quantitatively. The most important field of application of this instrument is the analysis of transient or rapidly alternating potentials. In this service it has a great advantage over the galvanometer type of instrument in that the inertia of the electron beam is negligible in comparison with that of any mechanical system, and it can follow oscillations up to frequencies of the order of 10 megacycles. Also, the beam is capable of motion in two dimensions and one of these axes may be used as a time axis. The instrument can be used in various ways. V, E_y, and E_z are all variable if desired and deflections can also be produced by external magnetic fields, but there is little work that cannot be done by the electrostatic deflection method in which E_y and E_z are the only variables. There are two general techniques: (a) that in which sinusoidal potentials are applied to both sets of deflecting plates, and (b) that in which a sawtooth wave or linear time axis is applied to one set and the potential to be analyzed applied to the other. These two methods will be briefly described.

The first method is particularly adapted to alternating-current phase and frequency measurement. Consider first two potential waves of the same frequency applied to the two sets of plates. The deflections are given by

$$\begin{aligned} y &= Y \cos \omega t \\ z &= Z \cos (\omega t - \varphi) \end{aligned} \tag{7.18}$$

where Y and Z are proportional to the peaks of the potential waves applied to the two plates. If there is no phase difference between the two, as would be the case if they were the drops across two series resistances, $y = (Y/Z)z$. Thus the path of the spot on the screen is a straight line making an angle $\tan^{-1} Y/Z$ with the z axis. A phase difference of π or a reversal of one set of terminals would also lead to a straight line having a negative slope of the same magnitude. In the case of a phase difference of $\pi/2$, as would be the case if the potentials were the drops across a resistance and a condenser in series, the equations become

$$y = Y \cos \omega t \qquad \text{and} \qquad z = Z \sin \omega t$$

Eliminating t to obtain the path of the spot

$$\frac{y^2}{Y^2} + \frac{z^2}{Z^2} = 1$$

This is the equation of an ellipse with its principal axes along the coordinate axes and with major and minor axes proportional to the amplitudes of the potential waves. If these amplitudes are the same, as would be the case for a series resistance and capacity if $R = 1/\omega C$, the path is a

circle with a radius proportional to the line current. In general for any arbitrary phase the path of the spot is obtained by eliminating t between Eqs. (7.18), yielding

$$\left(\frac{y}{Y}\right)^2 - 2\frac{yz}{YZ}\cos\varphi + \left(\frac{z}{Z}\right)^2 = \sin^2\varphi$$

This is also an ellipse but with inclined axes. If the amplitudes of the two waves are the same, the pattern resembles the left-hand one of Fig. 7.31, and from the above equation the major and minor axes are in the ratio $[(1 + \cos\varphi)/(1 - \cos\varphi)]^{1/2}$. In any case the pattern is bounded by a rectangle whose sides are proportional to the amplitudes of the waves. As the phase is varied, the pattern changes from one diagonal through a series of ellipses that sweep out the entire rectangle till it contracts to

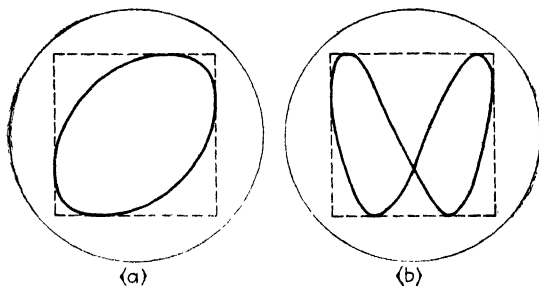



FIG. 7.31.—Cathode-ray-oscillograph patterns.

the other diagonal. If φ is a function of the time and the fluorescence persists for an appreciable time the entire rectangle will glow.

Consider next the case in which the potential waves of different frequencies are applied to the plate. In this case

$$y = Y\cos\omega_1 t \qquad \text{and} \qquad z = Z\cos(\omega_2 t - \varphi_2)$$

Here also the pattern is bounded by a rectangle whose sides are proportional to the amplitudes of the waves, but the pattern is not an ellipse. In fact, if the frequencies are incommensurate, the spot will sweep out the entire rectangle, for it is as if the phase difference between the waves were a function of the time. The most interesting cases, however, are those for which

$$n_1\omega_1 = n_2\omega_2 \tag{7.19}$$

where n_1 and n_2 are small integers. In this case the pattern is a closed curve and is known as a Lissajous' figure. A typical figure for $n_1 = 1$ and $n_2 = 2$ and equal amplitudes is shown at the right in Fig. 7.31. The analytical expressions for these curves are not very illuminating, but on looking at such a pattern it is easy to imagine that one is viewing a

transparent cylinder upon which is traced a sine wave, the axis of the wave being a perimeter of the cylinder. The illusion is enhanced if one of the phases changes slowly, *i.e.*, if $n_1\omega_1$ is not exactly equal to $n_2\omega_2$, for then the cylinder appears to rotate about its axis. The illusion is, of course, due to the fact that one of the sinusoidal motions, generally the slower, is envisaged as the projection of uniform circular motion. Consider that the sine wave is traced out on the cylindrical surface (as on a fluorescent screen) by a beam emerging from the cylindrical axis and rotating with an angular velocity ω_1. If the ends of the sine wave join after n_2 rotations, the time for n_2 rotations at an angular velocity ω_1 is the same as that for the description of, say, n_1 waves of period τ_2, or $2\pi n_2/\omega_2 = n_1\tau_2$, which is the same as Eq. (7.19). Thus the ratio of the frequencies of the waves is the ratio of the number of times the wave is wrapped around the cylinder to the number of wave lengths or maxima described. This is evidently the same as the ratio of the number of points of tangency with adjacent sides of the enclosing rectangle. This supplies a very convenient and accurate method of frequency measurement, for if one of the frequencies is known, the other is determined with equal accuracy from Eq. (7.19) when the pattern is stationary. If the pattern rotates slowly, the frequency of rotation can be added as a small correction term.

The other method of use of the oscillograph is to apply a saw-tooth wave to one of the plates and the potential to be analyzed to the other. If the wave shown in Fig. 7.32 is applied, say, to the z plates, the spot is drawn across the screen at an approximately constant rate for a time τ, then returns to its starting point in a negligible time. This occurs once per cycle, and if a sine wave with a period, say, τ/n is applied to the y plates, n sine waves will appear on the screen. The wave to be analyzed need not, of course, be sinusoidal. If the oscillograph is linear, *i.e.*, the displacement of the spot is proportional to the deflecting field, and if the slope of the saw-tooth wave is a straight line, the figure on the oscillograph screen will accurately reproduce the potential wave on the y plates n times across the screen. Thus recurrent wave forms can be analyzed at leisure and transient phenomena, due to the retentivity of the screen, can be seen for an appreciable time. This is the most satisfactory method of observing wave forms and it is available over a wide range of frequencies.

The linear time axis is obtained with a so-called sweep circuit which yields a voltage output of the form of Fig. 7.32. A condenser C is charged at a constant rate, *i.e.*, $V_c = it/C$, and when a certain limiting potential is reached, its discharge is rapidly brought about by the breakdown of a gas-discharge tube. The potential across the condenser is applied to one pair of deflecting plates. For the charging current i to be strictly

constant the condenser must be charged through some constant-current device such as the plate circuit of a pentode. However, an approximately linear axis can be obtained if the condenser is charged through an ordinary ohmic resistance if only a short portion of the exponential charging characteristic is used. A simple circuit is shown in Fig. 7.33. The condenser C is charged through the resistance R by the drop across the bleeder resistance and its potential is applied to the deflecting plates through the coupling condenser C'. The discharge is brought about by the gas-filled tube or thyratron (Sec. 8.4) which has the property of suddenly breaking down and providing a low-resistance path at a critical potential, say V'. This potential can be controlled by means of the grid bias, as shown in the figure. V' determines the height of the potential

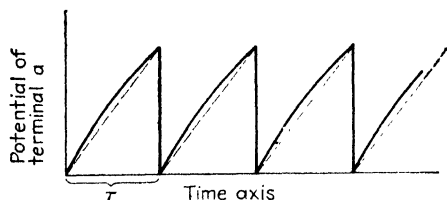



FIG. 7.32.—Saw-tooth wave for obtaining a linear time axis.

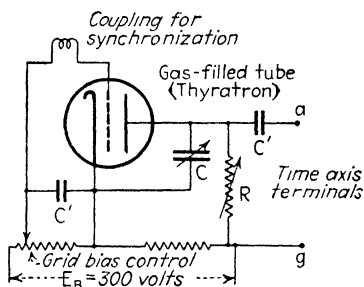



FIG. 7.33.—Sweep circuit for generating a saw-tooth wave.

wave and its period τ is proportional to the product RC. Thus the length of the time axis is controlled by the grid bias and the period is adjusted to an appropriate value by varying either R or C. It is frequently a great advantage to be able to synchronize this wave with the alternating-current potential which is to be analyzed so that the pattern will remain strictly stationary. This can be accomplished by some interconnection between the alternating-current circuit and the grid so that the latter receives small positive impulses with the period of the circuit under measurement. Thus the time of breakdown of the tube is controlled (within reasonable limits of the RC adjustment) by the period of the circuit being measured, and the sweep circuit will lock in with this period yielding a stationary pattern.

Problems

1. What electron current limited by space charge will flow per square centimeter of area between a plane cathode and anode 1 cm. apart in a vacuum if a potential difference of 1,000 volts exists between them?

2. Assuming that the total space-charge-limited current from a cathode to an anode is a function of the potential difference between them (V), the ratio of the charge of an electron to its mass (e/m), and possibly the linear dimensions of the apparatus (d), show that necessarily $i = A\kappa_0(e/m)^{1/2}V^{3/2}$, where A is a dimensionless constant.

3. Calculate the space-charge-limited current in amperes per centimeter length between a cylindrical cathode 1 mm. in diameter and a coaxial anode 2.72 mm. in diameter if the potential difference between them is 100 volts.

4. Assuming that a rectifier obeying Eq. (7.4) is in series with a resistance R and an emf. $\mathcal{E}$, calculate the efficiency of the circuit (power in resistance/total power in circuit). Assuming $d = 1$ cm. and an area of 1 cm.², and that $R = 10{,}000$ ohms, plot the efficiency as a function of i and the total emf. $\mathcal{E}$ as a function of i. From these parametric curves plot the efficiency as a function of $\mathcal{E}$.

5. Show that the equivalent input conductance (i_g/e_g) of a triode in the absence of a series grid resistance r_c is given by

$$k_g \frac{k_p(1 - \mu_p\mu_g) + k_l}{k_p + k_l}$$

Interpret μ_p and μ_g in terms of the plate and grid surfaces, $i_b(e_c, e_b)$ and $i_c(e_c, e_b)$.

6. Show that the equivalent output resistance (e_p/i_p) is given by

$$r_p \frac{r_g + r_c}{r_g + r_c(1 - \mu_p\mu_g)}$$

and that in the absence of i_g and other complicating effects $\mu_g = 0$ and the equivalent output resistance is r_p. What is the importance of this quantity and that of the preceding problem?

7. Draw the equivalent plate circuit corresponding to Eq. (7.11).

8. Show that the maximum power transfer from the plate battery to the load (alternating current) occurs when $r_p = r_l$.

9. Show that the voltage-amplification factor of a triode is given by

$$\frac{e_l}{e_g} = \frac{\mu_p}{1 + \dfrac{r_p}{r_l}}$$

Plot this expression as a function of r_l.

10. Given a plate battery of 300 volts, what is the proper value of the load resistance and the resultant value of the voltage-amplification factor if the tube characteristics are as follows: $E_b = 135$ volts, $I_b = 1.5$ ma., $r_p = 32{,}000$ ohms, and $s_p = 940$ micromhos.

11. Calculate the minimum value of the resistance in which an emf. to be amplified is generated for which each of the following tubes can be employed to advantage as an amplifier. Note the optimum value of the galvanometer resistance with and without the tube.

6-C-5	$\mu_p = 20$	$r_p = 10{,}000$ ohms
6-F-5	$= 100$	$= 65{,}000$
6-F-6	$= 200$	$= 75{,}000$

12. Show that the change in load current for a change in illumination dL of the photocell in Fig. 7.16 is given by

$$i_l = \frac{\mu_p R_0 a\, dL}{R_l + r_p}$$

where a is a constant of the cell (ratio of the change in cell current to the change in illumination).

13. Assuming that a battery of 25 volts is in series with the photocell of Fig. 7.16 and the resistance R_0 and that upon being illuminated it passes a current of 2×10^{-7}

amp., calculate the ratio of the incremental power in the load to the total power in the photocell circuit if $R_0 = 10$ megohms, $\mu_p = 20$, $r_p = R_l = 10{,}000$ ohms.

14. Show that in the case of the inclusion of a resistance r_f in series with the cathode of Fig. 7.4 the e.p.c. theorem becomes

$$i_p = \frac{\mu_p}{(1 + \mu_p)r_f + r_p + r_l} e_g'$$

where e_g' is the incremental potential between the lower end of the cathode resistor and the grid. Show that this is related to e_g (incremental potential between cathode and grid) by

$$e_g = \frac{r_f + r_p + r_l}{(1 + \mu_p)r_f + r_p + r_l} e_g'$$

15. Assuming that the voltmeter tube of Fig. 7.17 has the characteristic

$$i_b = I_b + ae_g + be_g^2$$

show that the meter deflection is proportional to the mean-square value of e_g.

16. Show that if the tubes in the wattmeter circuit of Fig. 7.18 have a predominantly parabolic characteristic, the deflection of the direct-current meter is proportional to the power consumed in the line circuit.

17. Show that if $e_0 = 0$, the vacuum-tube circuit of Fig. 7.4 can be considered as separate plate and grid circuits, each composed of a resistanceless current generator in parallel with the element and the external circuit resistance, *i.e.*, $s_p\mathbf{e}_g = (\kappa_p + \mathbf{Y}_e)\mathbf{e}_l$ and $s_g\mathbf{e}_p = -(\kappa_g + \mathbf{Y}_c)\mathbf{e}_G$.

18. What is the effective charging current drawn by a 1-μf. condenser at: (*a*) 60, (*b*) 1,000, (*c*) 100,000 cycles at $V_e = 10$ volts?

19. Calculate the series resistance that will account for the power loss in a condenser with an imperfect dielectric.

20. Given that the power factor ($\cos \varphi$) of a 1-μf. condenser is 0.95 at 1,000 cycles, calculate the equivalent shunt resistance. How does the power factor vary with the frequency?

21. It is observed that the impedance of an imperfect condenser is 100 ohms at 100 cycles and 11 ohms at 1,000 cycles. Calculate the power dissipated in the condenser for an applied $V_e = 100$ volts.

22. A resistance of 10 ohms and a condenser of 10 μf. are in series with the shunt combination of a 20-ohm resistance and a 5-μf. condenser in a 100-volt 60-cycle line. Calculate the peak potential across both condensers, the effective current in each branch, and the power factor of the circuit.

23. Obtain the solution [Eq. (7.14)] of Eq. (7.12′) by means of Eq. (*C*.6) of the Appendix, neglecting the transient term and assuming $V = V_0 \cos \omega t$.

24. Obtain the complete solution of Eq. (7.12′) by Eq. (*C*.6) if $V = V_0 \cos \omega t$ with the auxiliary condition $i = 0$ at $t = 0$.

25. If a complex potential wave is applied to a resistance R and capacity C in series, calculate the amplitudes and phases of the harmonics of the potential wave appearing across C.

26. An alternating-current ammeter in series with a condenser in a 100-volt 60-cycle line indicates 1.34 amp. What is the capacity of the condenser?

27. It is desired to measure the capacity of a condenser of approximately 0.01 μf. which is suspected of having a leakage resistance of about a megohm. What is the optimum frequency to be used and what should be the order of magnitude of the other bridge elements?

28. If a fluorescent screen is placed a distance a to the right of the circular diaphragm of Fig. 7.28, show that it must be maintained at a potential $4V$ in order that an image of the cathode should be formed upon it.

29. Assuming circular diaphragms and that the second is maintained at five times the potential of the first, calculate from Eq. (7.17) the ratio of the separations and from Eq. (7.16) the diameter of the emergent beam if the entrance diaphragm is 2 mm. in diameter and the beam emerges parallel.

30. Two triodes are connected together in such a way that the plate of the first is separated from the cathode of the second by a battery of emf. equal to the operating plate potential of the first, and the two grids are connected directly together. Considering this combination as a single triode with its input between the grid and cathode of tube 1 and its output between the cathode of 1 and plate of 2, show that the effective amplification constant of the combination is $\mu_2(\mu_1 + 1)$ and the transconductance is $\mu_2(\mu_1 + 1)/(r_{p1}\mu_2 + r_{p2})$.

CHAPTER VIII

ELECTRICAL CONDUCTION IN GASES

8.1. Elementary Phenomena.—The conduction of electricity in gases differs fundamentally from that in solids or liquids. The majority of the current is carried by electrons, and in this respect it resembles solid conduction. Positive ions are also present and free to move, but these heavy ions play a very different role than in liquid conduction. The principal phenomenon that differentiates gaseous conduction is the production of ions in the conducting medium by the current itself. This leads to a more complex situation than in other types of conduction, and the details of the behavior of various forms of discharges are not as yet completely understood. However, the individual elementary phenomena that contribute to the discharge mechanism have been very extensively investigated. On the basis of these studies it is possible to understand qualitatively the behavior of the various types of gas discharge, and many of the simpler ones can be accounted for on a quantitative basis. These elementary atomic phenomena will be considered under four heads: (*a*) the production of electrons and ions at electrode surfaces, (*b*) their production in the body of a gas, (*c*) their motion, and (*d*) their recombination in the gas and disappearance at the walls and electrodes.

Production of Electrons and Ions at Electrode Surfaces. 1. *Thermionic and Field Emission.*—This has been discussed in Sec. 6.5 and the presence of the gas has little effect upon it, except in so far as it cools the emitting surface or adsorbed layers, or chemical reactions alter the work function. In the case of a surface of low mechanical stability, such as that of certain oxides, the bombardment by high-energy ions may alter the surface and change its work function. For this reason the potential drop in tubes with specially prepared cathode surfaces must be kept low so that ion bombardment does not destroy the surface. In other types of tubes, such as the tungar rectifier, the cathode surface is that of a pure metal not injured by positive-ion bombardment. Here there is a real advantage in having a large potential drop, for the energy of the ions heats the cathode to incandescence and external heating can be dispensed with.

2. *Photoelectric Emission.*—This has also been discussed in Sec. 6.6. In discharges, however, its importance is not limited to the case of external illumination. The processes taking place both at surfaces and in the body of the gas are such as to be accompanied in general by the emission

of radiation. Most of this radiation is of too short a wave length to be visible. But because of its high frequency or the high energy of the photons it is all the more effective in liberating electrons from the electrodes. The discharge phenomena brought about depend on the nature of the gas. Slight traces of polyatomic gases, for instance, can have strong absorption for these photons over a wide frequency band with resultant dissociation of the polyatomic gas, a fact that is used in the construction of self-quenching Geiger counters (Sec. 8.6).

3. *Emission Due to Electron Bombardment.*—The secondary electron emission that was mentioned in Sec. 6.6 is also an important phenomenon in gas discharges. Carbon in general has a low secondary emission and hence is frequently used for tube elements when this effect is undesirable. Insulating surfaces may also emit electrons more rapidly than they are received. This results in a local positive charge which further increases electron bombardment and the spot may be raised to incandescence. It is a frequent cause of failure in high-potential discharge tubes. Usually the walls of a glass envelope become negatively charged for low electron velocities.

4. *Electron Emission Due to Positive-ion Bombardment.*—This is a much less efficient process than the previous one, but if positive ions with energies of several hundred volts impinge on a surface, they may liberate a certain number of electrons. The ratio of the number of secondary electrons emitted to the number of ions striking the surface is of the order of 0.01 to 0.1 in this energy range. It is a general characteristic of all discharge phenomena that the production of charged particles by positive ions is much less efficient than their production by electrons of the same energy.

5. *Electron Emission Due to Metastable Atoms.*—The phenomena described under the next heading result in the dislocation or disruption of the normal electron structures surrounding the atoms or molecules of the gas. In general, if none of the electrons are completely removed, the structure returns to normal by emitting its excess energy in the form of radiation. However, in certain instances, notably mercury atoms and those of the rare gases, the dislocation is of such a nature that the electron structure cannot return to normal by radiating and the atom carries this energy about with it for an appreciable time like a small bomb or compressed spring. Atoms in which this energy of excitation is retained are known as *metastable* atoms. When they come in contact with a surface, the energy can be released, and if it is transferred to an electron, the latter may be liberated. This phenomenon is closely related to emission due to positive-ion bombardment. When a positive ion approaches the surface, it can capture an electron from the metal before striking. In fact, at a distance of about 5×10^{-5} cm., the prob-

ability is almost unity of capturing an electron to form a neutral atom in an excited state with excitation $I' = I - \phi$, where I is the ionization potential of the atom and ϕ the work function of the metal. If $I > 2\phi$, it can transfer this excitation energy to an electron with resultant secondary emission. This can occur at distances of the order of 2×10^{-8} cm., and thus secondary emission will depend on the lifetime of the excited atom. This lifetime is large for rare gases and mercury. For polyatomic vapors, the metastable atom loses its excitation energy before it can transfer it to a surface electron, another factor in quenching a counting tube.

Formation of Electrons and Positive Ions in the Body of a Gas. 1. *Thermal Ionization.*—This is somewhat analogous to the thermal liberation of electrons or positive ions at a metal surface. As the temperature of a gas rises, the atoms or molecules move faster and faster till at an appreciable number of collisions the violence of the impact dislodges an electron from one of them and an electron and positive ion results. The number of ion pairs produced in this way is relatively small at low pressures, but in high-pressure discharges such as the carbon arc the phenomenon plays an important role.

2. *Photoionization.*—High-frequency radiation, corresponding to the far ultraviolet, is liberated when excited atoms in the discharge return to normal and it traverses the body of the gas. These photons may encounter other atoms that are capable of either totally or partially absorbing them. If the photon is of sufficient energy, the result of the absorption may be to eject an electron from the atom in much the same way as an electron is liberated photoelectrically from a surface. From the work-function point of view the value of φ for such an atomic process is of the order of 10 to 25 volts rather than from 1 to 6 as in the case of surfaces. Thus higher frequencies are necessary for photoionization than for surface photoelectric emission. This process plays an important part in producing ions uniformly throughout a discharge.

3. *Ionization by Electron Impact.*—This is by far the most important process for the production of ion and electron pairs in the body of the gas. Electrons, produced by the many sources of residual ionization in a gas such as heat, light, cosmic rays, etc., or liberated from the exposed surfaces, are accelerated by an applied electric field. As they move under the influence of the field, they gain energy at its expense and their velocity increases. At first, when they move slowly, they bounce elastically from any atom they encounter, but when their energy reaches a critical value, it can be transferred to the electron structure of an atom in the path and induce a so-called *excited state* in the atom. An atom is not able to accept an arbitrary amount of energy from an electron, but each type of atom has certain characteristic energies that it is able to

assimilate. Thus a mercury atom, for instance, can completely absorb the energy of an electron that has fallen through a potential difference of 4.85, 5.47, 6.73, etc., volts or it can accept this amount of energy from an electron that has more, leaving the electron with the remainder. These potentials are known as *excitation potentials*. For neon they are 16.6 and 18.5, and for helium they are 19.75, 20.55, 21.2, and 22.9 volts.

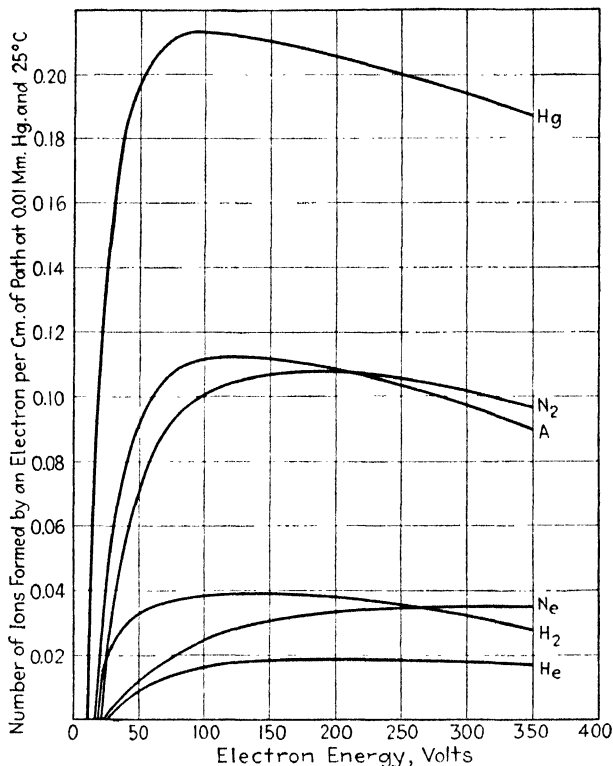



FIG. 8.1.—Formation of ion pairs by electron impact. (*Compton, Van Voorhis, & Langmuir.*)

After an atom has been excited in this way, it can in general return to normal by emitting light, and the color of the light is characteristic of the type of atom just as is the excitation potential. In the cases for which this method of ridding itself of the excess energy by radiation is not possible the atom remains in a metastable state.

If the electrons which are drawn through the gas are able to gain more than the excitation energy from the field, they may actually knock out an electron from the normal structure surrounding an atom. This also occurs only if the electron possesses more than a certain critical energy which in volts is known as the *ionization potential*. It is also a characteristic of the particular type of atom. In order to be able to

expel an electron from a mercury atom, the primary electron must have fallen freely through a potential difference of 10.39 volts. For neon this value is 21.47 and for helium 24.53 volts. For the alkali-metal vapors the ionization potential is very low, being between 4 and 6 volts; for other gases it lies between 10 and 25 volts. When the process takes place, two new electrical carriers are formed, the new electron and the residual positive ion. Also the probability of its occurrence is large so that it is of primary importance in discharge phenomena. The accompanying curves of Fig. 8.1 show the number of ion and electron pairs formed per centimeter path at a pressure of 0.01 mm. Hg by an electron of a given energy in the various gases. This number is seen to reach a maximum at an energy that is of the order of 10 times that at which the process is first able to take place. If the electron is going sufficiently rapidly, it can knock out more than one electron at a time. With an energy of several hundred volts it may eject as many as five electrons at once from a mercury atom. Such occurrences, of course, greatly increase the number of electrons available for carrying the discharge current.

4. *Ionization by Positive-ion Impact.*—Positive ions on moving through a gas may also break up atoms in their path if they have sufficient energy. But this energy must be larger than that of an ionizing electron by a factor of the order of 100 before the effect is detectable. Even then the ionization per centimeter path under the same conditions is generally smaller, though in the case of very high energy particles the probability of ionization is about the same for ions as for electrons of the same velocity. Hence the process is of less importance than the one previously described.

5. *Ionization by Metastable Atoms and Cumulative Ionization.*—Metastable atoms are analogous to small compressed springs and the energy that helium or neon metastables, for instance, carry about with them is more than sufficient to ionize many gas atoms or molecules such as those of mercury, hydrogen, or nitrogen. At a collision between a metastable atom and a normal atom of a different type there is a certain probability that the energy of excitation is transferred to the normal atom, and if it is in excess of the ionization energy, an electron is ejected from the normal atom and an ion pair is formed. This process is an important one, particularly in connection with small quantities of impurities in a discharge. Thus a trace of nitrogen in neon greatly facilitates the maintenance of a discharge. Neon metastables on colliding with nitrogen molecules produce free electrons and their contribution to the discharge current is vital under certain conditions. Finally it should be mentioned that these various discharge processes need not always take place in one stage. Thus a metastable atom produced at

one electron encounter may be completely ionized at a second by an electron possessed of the requisite energy (ionization energy—excitation energy). Such processes are only important, however, if long-lived metastable states exist.

Motion of Electrons and Ions. *The Velocity-distribution Law.*—The atoms or molecules of a gas are known to be in rapid random motion colliding with one another and with the walls of the containing vessel. The interatomic collisions provide a means of transfer of energy from one particle to another and the net result may be shown to be that a certain characteristic distribution of velocities exists among the atoms of the gas. This is such that the number of atoms per unit volume having the velocity v along any one of the three Cartesian axes lying in the small range dv is given by

$$nf(v)\,dv = n\left(\frac{m}{2\pi kT}\right)^{\frac{1}{2}} e^{-\frac{mv^2}{2kT}}\,dv$$

Here m is the mass of an atom and the equation is known as the *Maxwell-Boltzmann distribution law.* If the gas is composed of neutral atoms, positive ions, and electrons and if the latter do not recombine and are not subject to any external electrical forces, it might be expected that this distribution law would hold for all these different particles. These simplifying assumptions are generally not entirely justified, but they lead to a useful approximation. If in such an aggregate attention is fixed say on the electrons, it will be found that a distribution of velocities characterized by the above law exists among them. The temperature entering there is generally not the same as that characterizing the velocity distribution among the neutral atoms. In particular, if electric fields exist as in a discharge, the electrons will acquire much higher energies than the neutral particles and the value of T_e characterizing their distribution will be much higher than the ambient temperature. This is also true though to a lesser degree for the positive ions.

The number of particles crossing a unit area in unit time with a velocity component perpendicular to the area lying between v and $v + dv$ is given by $nvf(v)\,dv$. Integrating this expression over all values of v from zero to infinity yields the number of particles crossing a unit area per second in a positive sense. Calling this number n'

$$n' = n\left(\frac{kT}{2\pi m}\right)^{\frac{1}{2}} \tag{8.1}$$

If these particles are electrons, the random current density is given by

$$I_e = en'_e = 2.48 \times 10^{-14} n_e T_e^{\frac{1}{2}} \text{ amp. cm.}^{-2} \tag{8.2}$$

Here the numerical constant has been evaluated and n_e and T_e are the

number of electrons per cubic centimeter and the effective electron temperature, respectively. The distribution function can also be applied to particles subject to external force fields. Under these conditions, the quantity $mv^2/2$, representing the kinetic energy of the particle, must be replaced by the total energy, potential plus kinetic. Integrating over-all values of velocity (or kinetic energy),

$$n = n_0 e^{-\frac{U}{kt}} \tag{8.3}$$

where n is the density of particles per unit volume in the region of space where the potential energy is U and n_0 is the density at the point where U is taken as 0. If the potential in one region of a discharge is V_1 and in another V_2, then

$$\frac{n_2}{n_1} = e^{\frac{-e(V_2 - V_1)}{kt}}$$

gives the relative concentrations of electrons in these two regions.

The Mean Free Path.—The average distance traversed by a particle between collisions is known as the mean free path. It is a function of the density of the gas and the velocities and effective cross-sectional areas of the particles. Consider first a particle of effective radius r_1 which is projected through a swarm of other particles of radius r_2 that are moving so slowly they may be considered relatively at rest. The chance of the projectile striking one of the particles per unit length of its path is equal to the ratio of the collision cross-sectional areas per unit volume to a unit area. This is $n\pi(r_1 + r_2)^2$, where n is the number of particles per unit volume. The reciprocal of this quantity is the average distance traversed without colliding which is the mean free path λ. If the projectile is an electron r_1 is negligible, and

$$\lambda_e = \frac{1}{n\pi r^2}$$

where r is the effective radius of the gas atoms. In case the projectile is itself one of the gas atoms, the radii are the same and owing to the velocity distribution law, a factor $\sqrt{2}$ makes its appearance. In this case

$$\lambda = \frac{1}{4\sqrt{2}n\pi r^2}$$

Thus the mean free path of an electron is about $4\sqrt{2}$ times that of an atom of the gas. These expressions give correct orders of magnitude, but the effective radius r itself depends on the relative velocity and on the definition of what constitutes a collision. The concept of mean free path is useful, but it is not very sharp and the above expressions merely

relate it to the quantity r which is to be explained in terms of observational data. Assuming the perfect-gas law and an atomic radius of the order of 2×10^{-8} cm., the electron free path is given approximately by $T/10p$ in cgs. units. Thus at normal temperature and pressure λ_e is about 3×10^{-5} cm. or at a pressure of 0.1 mm. Hg it is about 3 mm.

Diffusion and Drift Velocity.—In addition to the perfectly random motion of the Boltzmann distribution law the charged particles will drift through the gas if concentration or potential gradients exist. Such gradients exist in discharges because of the nonuniform production and distribution of ions and because of the fields applied by the electrodes. The question of diffusion has already been mentioned in Sec. 6.2. There a diffusion coefficient D was defined. The net number n_1' of particles of type 1 diffusing across a unit area per unit time in the direction of the concentration gradient is given by

$$n_1' = -D_{1,2} \,\mathbf{grad}\, n_1$$

The second subscript of D indicates that the rate of diffusion is dependent on the other types of particles present, say particles of type 2. D is of the order of the product of the average velocity and the mean free path. If an electric field $\mathbf{E}$ exists, there is, of course, a force on a particle of charge e that is given by $e\mathbf{E}$. In the absence of other particles in its path it would experience a uniform acceleration. However, at collisions these obstructing particles deflect it and share its energy so that in general a terminal velocity is reached which is proportional to the accelerating field. The constant of proportionality between this terminal drift velocity and the field is the mobility u, *i.e.*, $\mathbf{v}_t = u\mathbf{E}$. The mobility is not strictly a constant independent of $\mathbf{E}$, but nevertheless it supplies a useful concept.

D and u are related in the following way: Consider a small region of gas in equilibrium under the influence of a pressure and potential gradient. The equilibrium implies equality between the body forces, or

$$-\mathbf{grad}\, p = ne\mathbf{E}$$

Also the net mass motion is the same. Assuming the perfect-gas law, the motion due to the concentration gradient is $-D \,\mathbf{grad}\, n = -\dfrac{D}{kT} \mathbf{grad}\, p$. That due to the field is $n\mathbf{v}_t = nu\mathbf{E}$. Equating these two and using the condition for the equality of the forces, it is seen that $u = eD/kT$. This is a very useful expression as it is not often that both D and u can each be measured experimentally. In the case of an ion at room temperature this expression shows that u is of the order of $40D$ in practical units.

Disappearance and Recombination of Electrons and Ions. *Disappearance at Surfaces.*—The great majority of electrons and ions in a discharge disappear at the various bounding surfaces. The net number that reach the electrodes constitute the external current, the others are merely neutralized at the walls. They may incidentally produce various of the other phenomena that have been previously discussed. A certain amount of heat is also generated. This is roughly measured by the number and kinetic energy of the particles striking the surface. In addition a certain amount appears as heat of neutralization and absorption analogous to the heat of condensation of a vapor. However, the total kinetic energy is not always delivered to the surface, but the ion may be reflected, either in its original state or neutralized, and carry with it a fraction of its original energy. The ratio $(E_i - E_r)/(E_i - E_s)$, where E_i and E_r are the initial and reflected energies, respectively, and E_s is the energy corresponding to the surface temperature, is known as the *accommodation coefficient.* For positive ions this is often a fairly small fraction, but few of them are reflected without contributing their heat of neutralization. Since the electron mass is much smaller than that of an ion Eq. (8.1) shows that the number of them crossing a unit area per unit time is much larger. Hence they diffuse to the walls more rapidly than the ions and if the walls are insulators, they acquire a net negative charge. Under ordinary discharge conditions this process continues until such a retarding field for the electrons is established that they and the positive ions reach the walls in equal numbers. In this equilibrium condition there is no net flow of current except to the electrodes.

Recombination in the Gas.—The loss of electrons and positive ions due to their meeting and recombining in the body of the gas is relatively small. If an electron and ion that approach one another closely are to adhere and form a neutral atom, there is excess energy that must be disposed of in some way. It is easy to see that in this type of two-body collision in which the additional energy eV_i (V_i = ionization potential) appears it is not possible to satisfy simultaneously the conditions of conservation of energy and momentum. Thus the energy cannot appear as kinetic energy of a single resultant particle. There are two other possibilities. In the first place a three-body collision can occur. Here the conservation equations can be satisfied, but except at very high pressures such an occurrence is very unlikely. There is a special type of three-body collision that is more important. Under favorable conditions slow electrons may become attached to certain types of molecules and form a negative ion. If this encounters a positive ion, the two may neutralize one another, the excess energy being distributed between the two neutral particles that result. This is undoubtedly an important

process and takes place very appreciably at high pressures. The effective cross section for positive- and negative- ion recombination may be as much as several thousand times the ordinary kinetic-theory value. The probability of this process and that of electron attachment depend markedly on the temperature and the type of molecule concerned. For instance, an electron would remain free and unattached about 200,000 times as long in CO as in Cl_2 (though in CO at normal temperature and pressure it would only remain free for the order of 0.001 sec.). Small traces of impurities, such as a few hundredths of a per cent of air in CO, reduce the average life of an electron by a factor of 10. However, important as this process is at high pressures, it is negligible in comparison with surface recombination in low-pressure discharges.

Another possibility for the recombination of an electron and a positive ion is that the excess energy appears as radiation when the neutral atom is formed. The light emitted at such a process has actually been observed, though it is but a very small fraction of the light emitted by a discharge tube. The great majority of this illumination is produced when excited atoms return to their normal states. This process of recombination appears to be a relatively unimportant one. Even in the most favorable case of a very slow electron the effective cross section for it is of the order of 0.01 times the kinetic-theory cross section. The process can therefore be neglected if there are any surfaces present or if the pressure is high enough for an appreciable formation of negative ions.[1]

8.2. Conduction in Gases at Low Current Density (Townsend Discharge).—Consider a region of gas that is contained between two large plane electrodes. If there is a small potential difference between the electrodes current will flow, and as this potential difference is increased, a critical value called the *breakdown potential*, V_b, will be reached at which the type of conduction taking place changes discontinuously into a glow, arc, or spark, depending on the particular conditions in the discharge and on the external circuit. In this section the preliminary conduction phenomena that occur below the value V_b will be described. Because of their much smaller mass the electrons move more rapidly in a field than do positive ions. As the current density is the product of the charge density and the velocity, all but about 1 per cent of the current is carried by electrons (neglecting the possibility of positive-ion-emitting surfaces). The electrons that carry the current at low potentials are those liberated thermionically or photoelectrically from the electrodes or those formed throughout the body of the gas by sources of residual ionization such as heat, light, cosmic rays, etc. The number of electrons formed in this way per cubic centimeter per second is of the order of 10

[1] COMPTON and LANGMUIR, *Rev. Med. Phys.*, **2**, 123 (1930).

under ordinary conditions. The electrons formed in these various ways are drawn to the anode. If the potential difference is increased till these reach the anode as fast as they are produced, the electrical characteristic flattens off as shown in Fig. 8.2. The characteristic up to this saturation current resembles that oi a high-vacuum thermionic tube or photocell.

If the potential difference is further increased to the point where some of the electrons are able to form additional electrons on colliding with atoms of the gas, the current rises above the saturation value. This increase is found to be exponential as would be expected from the following argument: Assume that at the existing pressure and applied voltage each electron is able to form α new ones per unit path length. Then, if there are n electrons drifting toward the anode per unit area per second at a distance x from the cathode, they will form $n\alpha\,dx$ new electrons in a distance dx. Hence the increase in the number of electrons over this distance is given by

FIG. 8.2.

$$dn = n\alpha\,dx$$

Integrating this expression with the condition that n_0 is the rate of emission of electrons per unit area of the cathode surface at $x = 0$ and assuming α constant

$$n = n_0 e^{\alpha x} \tag{8.4}$$

The electron current density reaching the anode, say at a distance d, is given by ne at $x = d$, or

$$i = en_0 e^{\alpha d} \tag{8.5}$$

This expression is found to agree with experiment over an increase in current by a factor of 20 or 30 as indicated in Fig. 8.3. This agreement indicates that the general concept is correct and that it is the only important process for this range of low currents. The value of α can be obtained from the semilogarithmic plot. The characteristic through this region is similar to that of the gas-filled photocell of Fig. 6.21.

If the potential difference between the electrodes is further increased, or if d is increased and the gradient is kept constant, the current eventually rises more rapidly than the exponential law would predict. Evidently other processes are brought into play that contribute to the production of electrons. The positive ions produced in the gas may eventually ionize neutral atoms in their path as they drift toward the cathode, or on reaching it they may liberate secondary electrons. Or the radiation emitted by excited atoms in the region may liberate photoelectrons from the cathode. All of these processes may make some con-

tribution, but as they all lead to the same form of characteristic, it is not possible to distinguish between them. The total number of positive ions that are produced in the gas and drift toward the cathode per unit area per second is equal to the difference of n at $x = d$ and n at $x = 0$ or $n_0(e^{\alpha d} - 1)$. The number of positive ions striking the cathode is also proportional to this quantity and so also is presumably the number

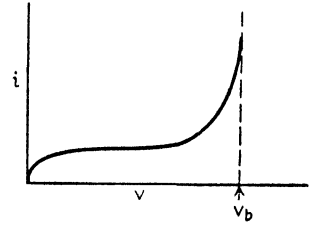
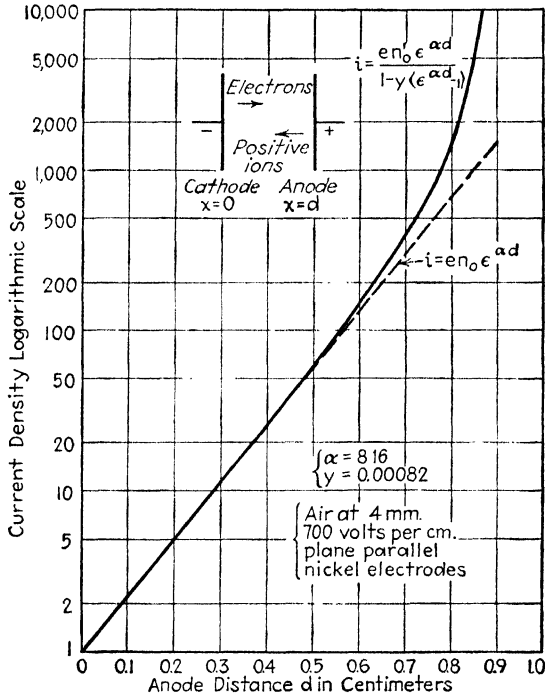



FIG. 8.3.—Curve representing the characteristics of the Townsend discharge.

of excited atoms formed in the region. As a consequence of any of these processes the number of additional electrons emitted from the cathode will be, say, $\gamma n_0(e^{\alpha d} - 1)$, where γ is a constant characteristic of the gas and the cathode surface. Thus, if n_0 is the total number of electrons emitted from the cathode and n'_0 is the number of primaries that are emitted in the absence of this process

$$n_0 = n'_0 + \gamma n_0(e^{\alpha d} - 1)$$

or

$$n_0 = \frac{n'_0}{(1 - \gamma(e^{\alpha d} - 1))}$$

And the total current to the anode which is given by Eq. (8.5) is

$$i = \frac{en'_0e^{\alpha d}}{1 - \gamma(e^{\alpha d} - 1)} \tag{8.6}$$

For the region of interest the second term in the denominator is less than unity so that as α increases due to an increased gradient or as d is increased, the denominator becomes smaller and the current larger. Thus the rise in current is more rapid than exponential, as indicated by the solid line in Fig. 8.3. This type of expression agrees remarkably well with experiment. The constant α depends on the type of gas, its pressure, and the potential gradient; for air at 4 mm. Hg and a gradient of 700 volts per centimeter, for instance, it is of the order of 8. The constant γ depends on the cathode surface as well. It is considerably smaller than α; under the previous conditions and with ordinary nickel electrodes it is of the order of 0.001.

Equation (8.6) is of particular interest because it shows that when $e^{\alpha d}$ increases to the value $(1 + \gamma)/\gamma$, the denominator vanishes and the current tends to become infinite. The value of V corresponding to this value of α is the breakdown potential V_b. Since $\gamma(e^{\alpha d} - 1)$ is the ratio of the emission induced by the discharge to the total cathode emission the breakdown condition, $1 = \gamma(e_b^{\alpha d} - 1)$, implies that every electron leaving the cathode gives rise to another one to take its place. The value of γ can be found either by fitting the theoretical expression to experimental points or from the breakdown condition; these two methods give values of the constant that are in excellent agreement.

It is found by experiment that α is proportional to the gas pressure and to a function of the potential gradient divided by the pressure. This would be expected on the following simple argument: The probability of an ionizing collision is a function of the energy of the electron (Fig. 8.1). The energy gained from the field is proportional to $E\lambda$, or V/pd since $E = V/d$ and λ is inversely proportional to p. Also the number of collisions per unit path is inversely proportional to λ or directly proportional to p. The product of these is the number of electrons formed per unit path or

$$\alpha = p\, f(V/pd)$$

Substituting this value of α in the breakdown condition shows that $pd\, f(V_b/pd) =$ const. or $V_b = F(pd)$. That is, the breakdown potential depends on the pressure and electrode separation only through the product pd. This is known as *Paschen's law*. It is a very useful relation, for if V_b is known for all pressures at a particular electrode separation, its value is known for any pressure and separation. The form of V_b as a function of pd depends of course on the gas and the nature of the cathode surface. Representative curves for air, hydrogen, and helium are shown in Fig. 8.4. These curves are for a particular cathode surface though the general form is typical. The area between the curve and the V and pd axes is the region of the Townsend discharge that has been discussed in this section. The lowest point of the curve represents

the minimum value of the potential at which a discharge can be maintained in the gas. It is approximately equal to the *normal cathode fall* which is the lowest voltage at which the later forms of discharge can be maintained between the electrodes.

The phenomena that occur for V greater than V_b will be discussed in subsequent sections. Since the discharge current tends to increase

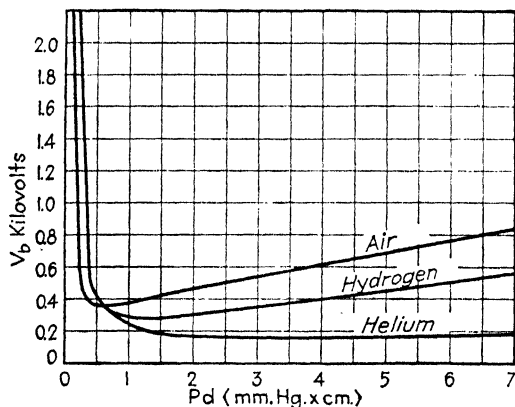



FIG. 8.4.—Typical breakdown-potential curves. (*Carr.*)

without limit at V_b, the over-all characteristic of the circuit will depend very much on the external resistance. Figure 8.5 represents a typical discharge characteristic. Let the external ohmic resistance be R. If the potential of the battery or generator is V_a, the current is normally given by i_1. But even before this corresponds to the value V_b, some disturbance may cause the operating point to shift from A to C (at B the dynamic resistance is negative, hence it is unstable). In order to trace out the entire characteristic without such a discontinuity it is necessary that R should be greater than the greatest negative slope of the discharge characteristic. Therefore in experimental investigations a high-potential source and large series resistance should be used. Often, however, the existence of later forms of the discharge so alters the cathode surface or temperature that the current changes with these variables as well and no proper characteristic of the i-V type exists. These later discharge forms will now be briefly considered.[1]

FIG. 8.5.—Graphical analysis of a discharge characteristic.

[1] THOMSON, J. J., and G. P. THOMSON, "Conduction of Electricity through Gases," Cambridge University Press, London, 1928; DARROW, "Electrical Phenomena in

8.3. Hot-cathode Low-pressure Discharge.—If the electron mean free path is considerably greater than the linear dimensions of the discharge tube, these later stages of the discharge will not develop. This is evident from the curves of Fig. 8.4. In this section low pressures will be considered as those for which the discharge develops but for which the free path is greater than about a thousandth of the tube dimensions. Above the breakdown potential V_b the discharge current increases rapidly. It continues to be carried almost exclusively by the electrons, but the positive ions, which are now produced in large quantities, begin to play an important role. Their effect is principally so to alter the potential distribution throughout the space that the anode effectively becomes very much closer to the cathode. In the case of a hot cathode from which the emission is limited by space charge

$$I_e = \frac{4\kappa_0}{9}\sqrt{\frac{2e}{m}}\frac{V^{3/2}}{d^2} \tag{7.4}$$

The effective value of d becomes much smaller as a positive-ion space charge develops and hence the electron current is increased. The value of V, which is approximately the breakdown potential, is in general not greatly altered. The increased current and enhanced rate of ion production are accompanied by radiation and possibly also metastable atoms. These additional agencies contribute to the uniformity of the discharge and may also contribute to the cathode emission. This latter effect is of major importance only in cold-cathode discharges.

The processes taking place may be visualized qualitatively by assuming that the electrons produce positive ions uniformly throughout the region between the electrodes. The electrons, being much lighter and effectively at a higher temperature, move more rapidly than do the positive ions. They are produced in equal numbers and have the same magnitude of charge, so from Eq. (8.1) the ratio of the number crossing unit area per unit time or the ratio of the random currents is

$$\frac{I_e}{I_p} = \frac{n'_e}{n'_p} = \left(\frac{T_e m_p}{T_p m_e}\right)^{1/2} \tag{8.7}$$

Hence the electrons diffuse more rapidly through the discharge and if a field exists, they are also drawn out more quickly. Thus, though they carry most of the current, if a potential gradient exists, they remain in the gas a much shorter time and their contribution to the space charge is negligible in comparison with that of the positive ions. If there is any field drawing electrons to the electrodes, an error of only about 1 per

Gases," The Williams & Wilkins Company, Baltimore, 1932; DRUYVESTEYN and PENNING, *Rev. Mod. Phys.*, **12**, 87 (1940).

cent will be made by neglecting them in space-charge calculations. Let s be the uniform rate at which ions are formed in the discharge and move toward the cathode. If $s = 0$, the potential distribution between the electrodes is a straight line, as shown in Fig. 8.6. Choosing the zero of coordinates at the anode, the rate of flow of positive-ion current past a point x_0 is given by

$$(q_v)_{x_0} v_{x_0} = I_{x_0} = \int_0^{x_0} se\,dx = sex_0$$

As the positive ions lose little energy in moving through the discharge, one may consider that an ion formed at x has the kinetic energy

$$\frac{1}{2} m_p v_{x_0}^2 = e(V_x - V_{x_0})$$

approximately, when it reaches x_0. The positive-ion space-charge density is then given by

$$(q_v)_{x_0} = \int_0^{x_0} \frac{se\,dx}{v_{x_0}} = s\left(\frac{me}{2}\right)^{\frac{1}{2}} \int_0^{x_0} \frac{dx}{(V_x - V_{x_0})^{\frac{1}{2}}} = -\kappa_0 \left(\frac{d^2 V}{dx^2}\right)_{x_0}$$

The last equality is due to Poisson's equation. A particular solution of the last equality is $V_x = V_c x^2/d^2$, where V_c is the cathode potential and d is the electrode separation, if

FIG. 8.6.—Building up of positive-ion space charge in a discharge.

$$(q_v)_{x_0} = \frac{-2\kappa_0 V_c}{d^2} = \frac{\pi}{2} s' de \left(\frac{m}{-2V_c e}\right)^{\frac{1}{2}}$$

Here s' is the particular value of s for which the equation is satisfied by the assumed value of V_x. The assumed solution represents a parabola with its maximum at the anode, as shown in Fig. 8.6. If the rate of production of ions increases beyond s', the equation continues to have a parabolic solution, but the maximum is no longer at the anode but moves out into the gas. Under these circumstances the field does not draw the electrons to the anode. They move toward the maximum of the parabola and are trapped in the body of the gas. They then counteract the effect of the positive-ion space charge and produce the flat plateau of the upper curve in Fig. 8.6. Neglecting the small potential gradient that produces an electron drift current toward the anode, the electron and positive-ion densities are nearly equal throughout this region. Electrons, of course, tend to diffuse out of this space to the walls and electrodes. If this

region of equal electronic and ionic densities is to be maintained, there must be a retarding potential at its boundaries to prevent most of the electrons from leaving it. As the electron energies are of the order of 5 to 10 volts, this must be about the height of the plateau potential above the walls. This potential assumes such a shape and value that electrons diffuse out of it at the same rate that they are formed and the concentration remains constant with the time. This region in which the electron and ion concentrations are high and the potential gradient is small is known as the *plasma*. It is bounded by regions of large gradient known as *sheaths*. These two regions are characteristic of all low-pressure discharges.

Assume that a plasma of this type is maintained as would for instance be the case for a current density of about 25 ma. per square centimeter and an applied potential of about 40 volts between a hot cathode and an anode at room temperature in mercury vapor. This region is a swarm of neutral atoms, positive ions, and electrons. Though the ionization is intense, only about 0.1 per cent of the particles are ions or electrons; by far the greater proportion are always neutral molecules. Thus, if the number of mercury atoms per cubic centimeter is about 10^{13}, the density of ions is of the order of 10^{10} for a representative discharge. The ions are formed by the various processes that were described in a previous section. Most of them result from electron impact, the original electrons gaining their energy from the rise in potential between the cathode and the plasma. The electrons and ions diffuse about in the plasma, the former several hundred times as rapidly as the latter. At the bounding sheaths the slower electrons are turned back into the plasma and only the fastest ones can move against the potential gradient and reach the walls or anode. These maintain the negative charge on the walls. The positive ions that diffuse into a sheath are drawn from the plasma and neutralized at the walls. Equations (8.2) and (8.7) give the random electron and positive-ion current densities. As has been mentioned previously, I_e is much greater than I_p both because m_e is less than m_p and because T_e is greater than T_p. The latter is due to the fact that at an ionizing collision the excess energy is largely distributed between the two resultant electrons and little kinetic energy is imparted to the ion. The Maxwell-Boltzmann distribution that is found experimentally to exist among the velocities of the particles is presumably brought about by high-frequency oscillations in the plasma. At the boundaries of the plasma T_p is found to be approximately $T_e/2$, but in the interior it is probably very much smaller. The experimental values of T_e are of the order of 10^4 or 10^5 °K.; this corresponds to an energy of from 1 to 10 volts and is of course much higher than the temperature of the ambient gas. The electrical characteristic of the plasma has a negative slope.

An increase in current density results in the production of so many more new electrons that the potential gradient necessary to maintain the same drift current actually falls. The characteristic of the sheath regions is positive, and the over-all dynamic characteristic may be either positive or negative, depending on cathode emission, type of discharge, etc.

Since most of the electrons are reflected back from the bounding sheaths, they are regions of positive-ion space charge. These ions emerge from the plasma and are drawn to the walls much as electrons are drawn from a hot-cathode surface. In fact, the space-charge equation (7.4) can be applied immediately to these sheaths. Here m is now the mass of an ion, V is the sheath drop, and d is its thickness. For a mercury atom, for instance, $(m_p/m_e)^{1/2} = 607$ and the positive-ion current through the sheath is

$$I_p = 3.83 \times 10^{-9} \frac{V^{3/2}}{d^2} \text{ amp./area}$$

This is smaller by the factor 607 than the electron current that would exist under reversed potential conditions. At surfaces to which there is no net current flow this must be the same as the electron current that diffuses out of the plasma against the field in the sheath. The maximum value of I_p for which this is applicable is the random positive-ion current in the plasma. From the previous equations this is $4.09 \times 10^{-17} n_p T_p^{1/2}$ amp. per square centimeter for mercury ions. It is of the order of microamperes per square centimeter under ordinary conditions. For an I_p less than this value the thickness of the sheath is determined in terms of V and I_p by the space-charge equation. If the ionization in the plasma is increased, say by increasing the discharge current, I_p increases, and as the sheath drop remains practically constant, its thickness decreases.

A valuable method of investigating discharge phenomena is to insert an auxiliary electrode in the plasma and measure the current to it as a function of its potential with respect to the anode. The electrical characteristic of such a probe in a mercury discharge is shown in Fig. 8.7. Below about -18 volts there is a small net positive current due to the collection of more positive ions than electrons. In the region from -20 to -12 volts the rise in negative current is quite accurately exponential. Here more and more electrons reach the probe against the retarding field. The form of the curve is given by Eq. (8.3), where n_1 is proportional to the random electron current in the plasma, n_2 is proportional to the probe current, and V' is the potential of the probe with respect to the plasma. The fact that the current is accurately described by this equation shows that the electron velocities are distributed in accordance with the Maxwell-Boltzmann equation and the slope of a semilogarithmic plot yields the effective electron temperature T_e. Below about -18 volts a positive-ion space-charge sheath exists about the probe, above

this value an electron sheath begins to form, and is completed at about −12 volts. Above this value the electron-space-charge equation applies. Above about −5 volts the electrons are drawn to the probe with sufficient energy to form new ion pairs in the sheath, these neutralize the electron space charge and the current rises rapidly. The probe then becomes an auxiliary anode in the discharge.

Figure 8.8 illustrates the potential distribution in the various types of discharge sheaths. The upper curve represents a positive-space-

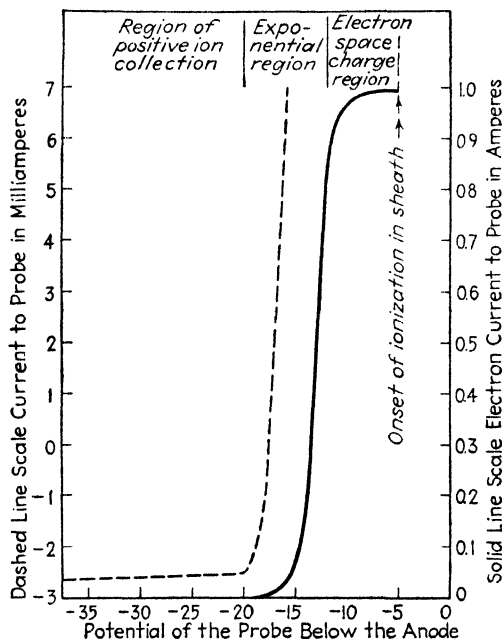



FIG. 8.7.—Probe characteristic in a mercury-vapor arc.

charge sheath such as exists above an insulating surface or a negative electrode. The lower curve is an electron sheath as found above a hot cathode in a high vacuum or between a plasma and an electrode sufficiently positive to collect a large number of electrons. Above a hot cathode in a discharge a double sheath exists. There is an electron sheath immediately above the cathode, followed by a positive-ion sheath between it and the plasma. If both these sheaths are completely formed, *i.e.*, if the ionization in the plasma and the cathode emission are both sufficiently copious, the potential distribution is symmetrical about the center. Each half resembles quite closely the distribution in an ordinary single sheath. The electron current from the cathode through such a completely formed double sheath is enhanced by the space-charge neutralization of the positive ions moving toward it and is equal to

1.86 times the electron current that would flow between the cathode and an anode at a distance apart equal to the sheath thickness. Thus the current through a double sheath is less than twice that through a single one of the same thickness and potential drop. When a plasma is formed in a hot-cathode discharge, say by decreasing the external resistance, this double sheath forms above the cathode and a single one above the anode. As the external resistance is further decreased, the discharge current and plasma ionization increase, the plasma grows, and the sheaths diminish in thickness. Here the slope of the characteristic is generally negative, which is typical of the plasma. At still smaller values of the external resistance the current becomes limited by cathode emission; the

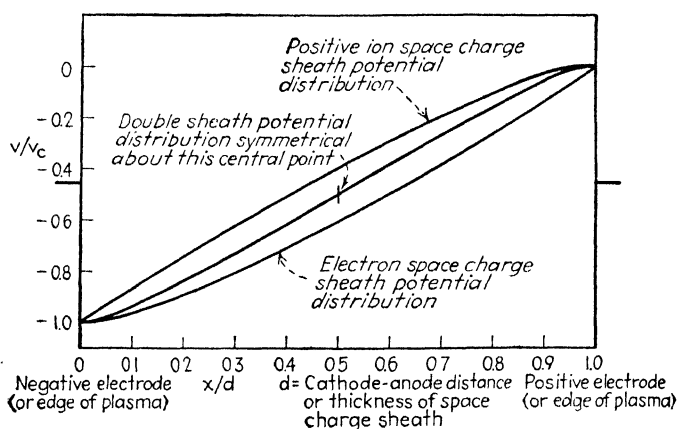



FIG. 8.8.—Potential distribution in typical discharge sheaths.

double sheath ceases to be symmetrical owing to the insufficient supply of electrons. The current rises less rapidly and the over-all characteristic of the discharge exhibits a positive slope. Furthermore, the geometry of the discharge tube affects the characteristic. If it is such that the area of the plasma is large in comparison with its volume, ions and electrons will be lost more rapidly and a greater potential gradient must exist to maintain the plasma. Also, if the supply of electrons is limited for any reason, the potential may rise from the plasma to the anode instead of falling, as shown in Fig. 8.8. Both these phenomena are associated with a positive slope of the characteristic.[1]

8.4. Practical Hot-cathode-discharge Devices.—Discharges of this type in mercury vapor or the rare gases are used for three general purposes: (*a*) lighting, (*b*) rectification, (*c*) control. When designed for illumination, the envelope is generally a tube an inch or so in diameter and several feet long. At one end is an anode (two if it is intended for

[1] COMPTON and LANGMUIR, *Rev. Mod. Phys.*, **3**, 191 (1931); LANGMUIR, *Gen. Elec. Rev.*, **38**, 452, 514 (1935).

alternating-current operation), and in a small bulb at the other is the cathode, which is generally an indirectly heated nickel surface coated with a mixture of barium and strontium oxides. Because the plasma is long and thin, the over-all voltage drop is large and the slope of the characteristic is positive. The current density is large, leading to intense ionization and excitation in the plasma. The light emitted when these atoms fall to states of lower excitation is the chief source of illumination. The light from a mercury discharge is blue-white, that from helium has a yellowish shade, and neon is characterized by an intense red. If the envelope is of glass, wave lengths shorter than about 0.3 μ do not emerge; to obtain ultraviolet radiation the tube must be of quartz. If two anodes are present, the tube can be operated on alternating current from a center-tapped transformer. Cold-cathode discharges and high-pressure arcs are also used for illumination; the former are generally used in low-power displays and signs, while the latter provide the most intense and concentrated sources of general illumination.

Gas and vapor rectifying tubes occur in various forms. The low ionization potential of mercury vapor makes it particularly well adapted to this purpose. In moderate power tubes the cathode is generally an indirectly heated surface with a low work function surrounded by heat-conserving shields. The envelope is roughly spherical in order to decrease the relative area of the plasma through which the losses occur. The form of the anode is not important. These rectifiers are much more efficient than the high-vacuum type because of the low potential drop and the smaller power required to maintain the cathode temperature. The cathode power and rated current for typical tubes of these two types are given below:

	Cathode power, kw.	Rated space current, amp.
High vacuum....	6.8	10
Vapor discharge......	0.32	72

As the potential rises across the elements in the proper sense, the current is negligible till the tube breaks down. Then as the plasma forms, the current rises and the tube drop falls till the region of emission saturation is reached, when the tube drop again rises. This is shown at the right in Fig. 8.9. The normal operating current is that for which the potential drop is lowest, namely, 2 amp. in the case of this tube. If higher currents are drawn, the discharge current is emission-limited and positive ions falling through the larger cathode drop are apt to damage the surface. In the case of a vapor discharge the pressure and mean free path are functions of the temperature. Hence the electrical characteristics

will depend on this parameter. The dashed line in Fig. 8.9 gives the tube drop for the normal current as a function of the ambient temperature. If the potential is applied to the tube in the reverse sense, no discharge ensues until the potential is sufficiently high to induce electron emission from the plate and establish a cold-cathode discharge. This is known as the *flashback potential* and it limits the inverse peak voltage that can be applied to the tube in a rectifier circuit. This flashback voltage is also a function of the temperature and is shown by the solid line in the figure.

This type of mercury-vapor rectifier has received the General Electric Company trade name of *phanotron*. For low-power rectification work

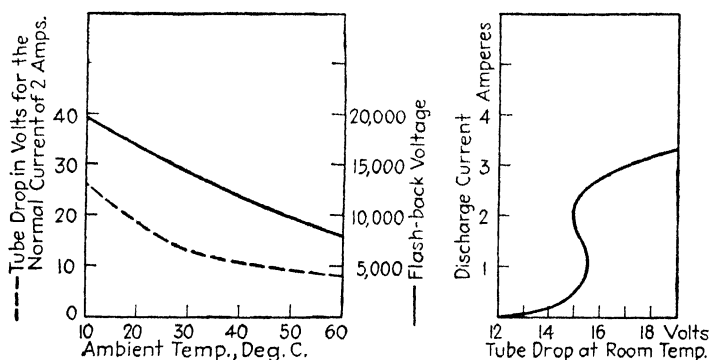



FIG. 8.9.—Temperature and electrical characteristics of the General Electric FG-32 phanotron (hot-cathode mercury-vapor rectifier).

argon-filled tubes with spiral tungsten filaments run at a high temperature are also used. These are known as *tungar* rectifiers. The electron emission from the tungsten is large and the high pressure of argon (about 5 cm. Hg) inhibits evaporation. The currents that can be carried are not great and the flashback potential is low (of the order of 200 volts). Once the discharge has started, the filament or any protuberance serving as a cathode can be maintained at incandescence by positive-ion bombardment. Thus the power necessary to maintain the cathode temperature can be supplied by the discharge itself if desired. For high-current work the mercury-vapor discharge above a mercury-pool cathode is frequently used. The discharge can be initiated by applying an instantaneous high potential, by running the mercury pool in contact with both electrodes, or by a high-resistance rod between them that causes a local discharge at the point of contact with the mercury and ignites the main discharge (*ignitron*[1]). Once started, the discharge is maintained by emission from the mercury cathode. The cathode mechanism is presumably similar to that in a high-pressure mercury

[1] Westinghouse Electric & Manufacturing Company trade name.

arc. Electrons are emitted by heat, radiation, and ion bombardment. The current densities may be very large, rising to as much as 5,000 amp. per square centimeter. As the cathode drop is of the order of 10 volts, the sheath above it is only about 10^{-4} cm. thick. Hence the gradient in it is enormous (10^5 volts per centimeter) and may reach such values above small ripples and protuberances as actually to pull conduction electrons out of the relatively cold surface. The characteristic of such a discharge is generally a rising one, indicating that the cathode release mechanism is more important than the plasma characteristic. If both electrodes are mercury pools, there is of course no rectifying action, for then both can serve as cathodes as in the high-pressure mercury arc. Such tubes are also used as spark gaps for the generation of damped oscillations in resonant circuits.

Gas or vapor discharges can be used for power-control purposes if one or more grids are inserted between the cathode and anode. The action of the grid is quite different from that in a high-vacuum tube, for once the discharge starts, the presence of the grid and its potential are immaterial. After breakdown it acts merely as a probe immersed in the plasma, and is isolated from it by one type of sheath or another, depending on its potential. It has then a very secondary effect on the anode current. However, the potential of the grid may control very accurately the anode potential at which the tube breaks down. Such grid-controlled discharges are known as *thyratrons*. By means of them a large amount of power in the anode circuit can be controlled by a very small amount in the grid circuit. As in the case of the simple phanotron, there is a maximum allowable or rated discharge current which implies an upper limit to the allowable anode potential after breakdown. The most useful characteristic curves of a thyratron are those giving the grid and anode potentials just before breakdown occurs. In the case of a vapor discharge these depend on the ambient temperature. The solid curves in Fig. 8.10 are those for the mercury-vapor thyratron FG-57 for three different temperatures in the normal range. For the type of grid design in this tube the breakdown occurs for negative values of the grid potential. For other grid designs such as those in the FG-33 and FG-67 the breakdown occurs for allowable anode potentials with a positive grid. The positive-grid tube, of course, draws more grid power as electrons reach it from the cathode, even before the discharge starts. And after breakdown there must be a protective series resistance to lower its potential and limit the current that it draws. However, it has the advantage that it remains in the nonconducting state for no grid excitation, *i.e.*, zero grid potential. The negative-grid tube draws much less current and hence requires a smaller power in the grid circuit. However, there must be negative grid excitation to prevent breakdown.

If the grid is simply disconnected, enough positive ions may reach it to set the tube off.

The random electron and ion current reaching the grid can be reduced by making it of very fine wires with the negative-grid design. This increases the direct-current input resistance. Also it may be screened by a second grid from the other tube elements. This increases the alternating-current input impedance and decreases the reaction of the anode potential upon that of the grid. Such a screen-grid thyratron is the FG-95 which is also argon-filled to decrease the effect of tem-

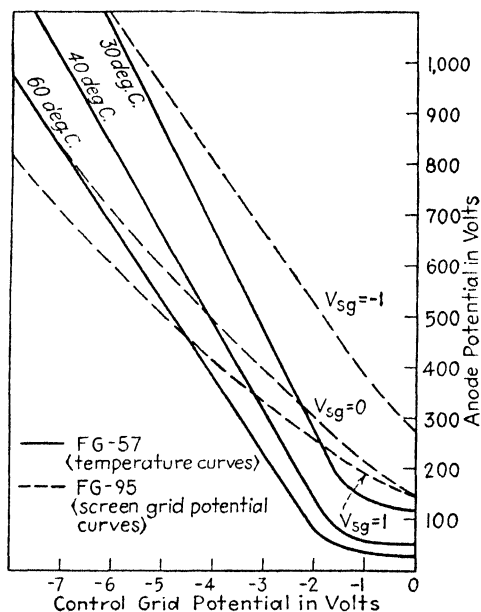



FIG. 8.10.—Thyratron breakdown curves.

perature fluctuations. This is a low-power control tube, but it has a very high input impedance (the random positive-ion current to the control grid before breakdown is of the order of 10^{-9} amp.). The screen potential is an additional parameter controlling breakdown. The breakdown characteristics for typical values of this parameter are shown by the dashed curves of Fig. 8.10. It is obvious that in all thyratron circuits the load resistance must be sufficiently great to limit the anode current to its rated value after breakdown occurs.

Once the tube breaks down, the grid loses control and the anode potential must drop below the value necessary to maintain the discharge before control is again achieved. Hence for direct-current operation some device must be used to lower the anode potential and reset the thyratron if the grid is ever to regain control of the anode current. In alter-

nating-current operation the anode potential falls below zero during the negative half of the cycle and grid control can then be reestablished.

Figure 8.11 illustrates the action of a thyratron in an alternating-current circuit. The positive loops of the anode-potential wave are solid and the negative ones are dashed. The solid curve below the positive-anode loops is to the grid scale and represents the grid potential necessary to prevent breakdown for the corresponding anode potential. It is obtained from the characteristic at 40° in Fig. 8.10. The dashed curve represents a potential wave applied to the grid. The discharge starts once each cycle at the point a when the grid potential intersects the breakdown curve. It is extinguished at b where the anode potential

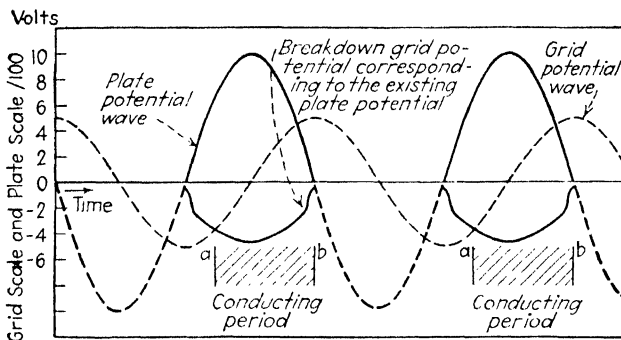



FIG. 8.11.—Illustration of the conduction characteristic of a thyratron in an alternating-current circuit as a function of the amplitude and phase of the grid potential.

drops to a value insufficient to maintain the discharge (this is practically the intersection of the anode-potential wave and the time axis). Thus the tube conducts for the time $a - b$ during each cycle. The anode current is limited by the external resistance and exhibits a very distorted wave. A variation in either the amplitude or phase of the grid wave will affect the average anode current. Both these methods are used in control circuits. The upper curve of Fig. 8.12 gives the average anode current as a function of the amplitude of the grid-potential wave at 180° phase difference (grid negative when anode is positive). Owing to the shape of the breakdown curve the current cannot be reduced continuously to zero by this method. The lower curve gives the average current as a function of the phase of the grid wave for an amplitude such as that shown in Fig. 8.11. With this type of control the current can be reduced to zero continuously. These curves can be obtained by altering the amplitude or phase of the grid wave of Fig. 8.11.

Figure 8.13 illustrates the use of a phototube and thyratron for the control of load power by illumination. A screen-grid type is used to match the high resistance of the photocell. A small transformer heats the cathode and supplies the photocell potential. The screen supply

that comes from a potentiometer across the power line can be used for auxiliary control. When the cell is illuminated, electrons can flow from the grid during the positive-anode cycle and the thyratron conducts; when the tube is not illuminated, the grid is effectively insulated and the tube does not conduct if the screen is sufficiently negative. Figure 8.14 represents a full-wave-rectifier circuit in which the fraction of the cycle

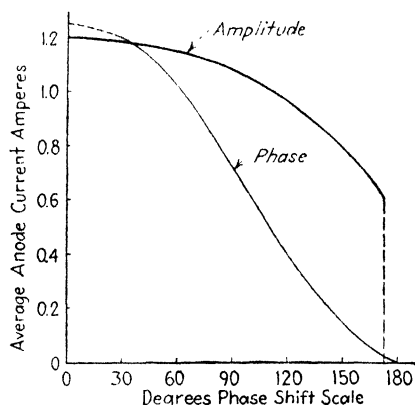



FIG. 8.12.—Variation of the mean anode current through a thyratron as a function of the amplitude of the grid-potential wave or of its phase relative to the plate-potential wave.

that the tubes conduct is controlled by the phase of the grid potential. The anode circuit is the same as that of Fig. 5.18a and need not be further discussed. The grid potential is obtained from a winding on the power transformer and applied symmetrically to the tubes through the small grid transformer. The resistances r_g are merely to limit the grid current to a small value after breakdown. The circuit containing R and C is

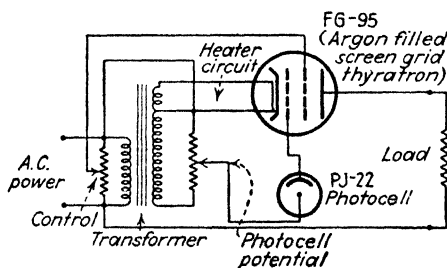



FIG. 8.13.—Photoelectric control of an alternating current with a thyratron.

for shifting the phase of the grid potential with respect to that of the anode. As R is varied from 0 to a very large value, the amplitude of the grid wave remains constant, but its phase is changed through 180°. This circuit draws very little power, but the adjustment of R controls the load current from zero to its maximum value. The action of the tubes is entirely symmetrical and the control mechanism in each is illustrated

by the curves of Fig. 8.11. This circuit is very useful in supplying a variable pulsating current from an alternating-current source. The output current resembles that of a full-wave rectifier (Fig. 5.18) when both tubes conduct an entire half cycle. As the period of conduction is decreased, the loops are cut off vertically from the left until they entirely disappear for a 180° phase difference between the grid and anode waves.[1]

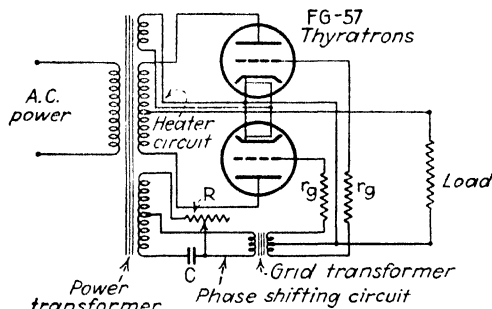



FIG. 8.14.—Full-wave phase-control rectifier.

8.5. Cold-cathode Discharges.—The cold-cathode discharge differs from those that have been discussed only in the mechanism of electron emission by the cathode. Before the discharge can develop from the Townsend stage to that characterized by the plasma, some of the activities of the earlier stage must produce a fairly copious cathode emission to take the place of the thermionic electrons in the hot-cathode discharge. When the positive ions, excited atoms, radiation, etc., which are produced by the Townsend discharge, appear in sufficient quantities, a cataclysmic effect due to the greatly enhanced cathode emission takes place, the current density rises above the value necessary to maintain the plasma, and the typical high current discharge sets in. The lowest voltage across the tube that can support this discharge is known as the *normal cathode fall*, so called because most of the potential drop occurs at the cathode. Referring to Fig. 8.4, if this occurs for a discharge in air at such a pressure that, say, $pd = 3$, the plasma extends from the anode to within such a distance of the cathode that the thickness of the cathode sheath times the pressure is approximately equal to 0.5 where the minimum of the breakdown curve occurs. Unless there is some change in the cathode surface, this later stage of the discharge does not develop below the pressure determined by this minimum point. The normal cathode fall is, of course, dependent on the gas and the nature of the cathode surface. Surfaces of low work function generally exhibit low cathode falls. Thus the value of this quantity in neon with a calcium

[1] MCARTHUR, "Electronics and Electron Tubes," John Wiley & Sons, Inc., New York, 1936; GULLIKSEN and VEDDER, "Industrial Electronics," John Wiley & Sons, Inc., New York, 1935.

or magnesium cathode is of the order of 90 volts. For air or hydrogen between nickel electrodes it is 200 or 300 volts. The current may be raised considerably above the minimum value at which the discharge can be maintained with little change in the potential drop across the tube. This is particularly true of the so-called *grid-glow tubes* having a large cathode surface. The current may increase as a result of increased actively emitting area if the external resistance is reduced. This involves no change in the electron-release process. However, any further increase in current after the total cathode surface is emitting requires an increase in specific emissivity. This requires a greater drop in potential in the cathode sheath, known as *abnormal cathode fall*, and the slope of the characteristic under these conditions is positive.

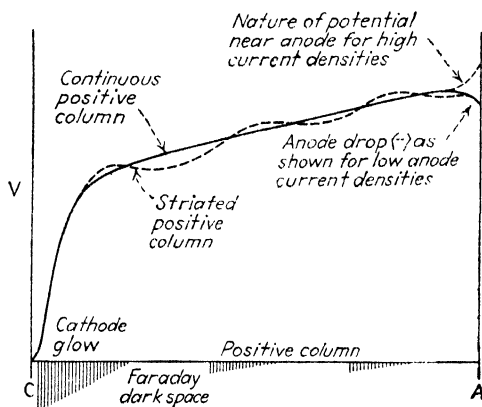



FIG. 8.15.—Potential distribution in a typical long discharge.

The dark space between the cathode and the neighboring plasma surface is known as the *Crookes dark space*, and the brilliant plasma edge where cathode electrons first produce excited atoms is called the *negative glow*. Much of the cathode fall occurs in this region, so it is somewhat different from the ordinary plasma. The ionization is so intense that the random electron current is large in comparison with that necessary to maintain a plasma; therefore few ions or excited atoms are produced in the subsequent region and it is relatively dark. This is known as the *Faraday dark space*. Electrons diffusing to the walls reduce the random current, and a field exists in this region in order that more electrons shall be produced to maintain the discharge. Hence this space is again followed by the typical glowing plasma, and from here to the anode it is generally known as the *positive column*. The periodic increase and decrease in electron density may continue all the way to the anode, yielding a striated discharge. The potential distribution is illustrated by the dashed curve of Fig. 8.15. The spatial

demarkation of the striae may be very sharp at high pressures, but as the pressure is decreased, they become less well defined owing to the random nature of the free paths, and they may fade into one another and be indistinguishable. Also, oscillations may exist that render them invisible except for stroboscopic observation. If discharge products such as radiation, metastable atoms, etc., are effective and diffuse rapidly from the place where they are formed, the positive column becomes a typical uniform plasma. Regions of nonuniform illumination may also be produced by constrictions in the envelope. In these the current density is high, causing intense illumination, but at either end where the tube widens out the current density decreases, enough electrons diffuse into these regions to carry the discharge current so few new ones are formed, and these regions are relatively dark.

The appearance of the discharge varies markedly with the type of gas, the pressure, the shape of the envelope, and the current density. The effects of these variables cannot be described in detail, but they are implied in the foregoing discussion. The volume of the plasma may be very strictly limited. Thus if the cathode emission is limited to a lime spot on a hot-metal surface and the pressure is moderately low, electrons are emitted normally and form a narrow pencil of luminous plasma down the tube. The positive-ion space charge effectively focuses the electron beam much as a system of diaphragms would. The electron trajectories may thus be made visible and it supplies a useful method of studying and demonstrating the effects on them of electric and magnetic fields. Certain types of cathode-ray oscillograph tubes employ this method of focusing, though the presence of the gas makes them less generally useful than the high-vacuum type. Also, if the cathode is a hollow cylinder the sheath may shrink at high-current densities till the plasma is almost completely confined to the interior of the cylinder. Electrons that would form an exterior plasma are rapidly lost to the walls, whereas those inside can only escape by diffusion out of the open ends of the cylinder. At very high currents the cathode temperature may rise to the melting point. If the current is maintained slightly below this value, the illumination from the cathode is very intense. Such a discharge in the rare gases is known as a *Schuler discharge* and because of evaporation the light is more characteristic of the cathode material than of the gas.

In all direct-current discharges a cathode and anode are necessary, but in high-frequency alternating-current fields a discharge may be maintained without any electrodes. It is known as an *electrodeless discharge,* and the plasma is analogous to an isolated positive column stria. For high-frequency oscillations electrons move back and forth through the tube maintaining the plasma without ever reaching the walls. Low-

pressure tubes that will sustain this type of discharge are very useful in the investigation of high-frequency circuits, for they will glow brightly in the presence of intense radio-frequency fields.

8.6. Relatively High-pressure Discharges.—At higher pressures where the mean free paths are shorter the breakdown phenomena assume somewhat different forms. The fundamental atomic processes taking place are undoubtedly the same, but the discharges are generally more localized and intense. For electrodes that are close together in comparison with their linear dimensions or radii of curvature the discharge takes the form of the familiar *spark* if the current is limited to fairly small values by an external resistance. This is a transient discharge localized in a jagged filamentary path. It is highly unstable, and if the emf. in the circuit is maintained, separate sparks follow one another in rapid succession. They follow separate paths and each spark appears to be independent of the rest. They are of such a transitory nature that investigation is extremely difficult and the detailed spark-discharge mechanism is not well understood. The potential at which a spark takes place between two electrodes depends to a certain extent on the pressure, temperature, humidity, etc., of the air. However, the effects of these parameters are not very large and the inception of spark breakdown can be used for rough measurement of high potentials. Tables of the sparking potential for various electrode forms will be found in the literature. The critical surface field for breakdown is given by the following empirical expression

$$E_c = A\rho\left(1 + \frac{B}{\sqrt{\rho R}}\right)$$

where R is the radius of curvature of the electrode, ρ is the density of the gas, and A and B are constants. For air at 25°C. $A\rho$ is approximately equal to 30 kv. per centimeter, and $B/\sqrt{\rho}$ is approximately 0.3.[1]

If any of the electrode radii of curvature are small in comparison with their separation, the spark breakdown is preceded by tentacular brush or treelike discharge from the regions of greatest field. This is known as *corona,* and it represents a stable form with a positive dynamic characteristic. It is commonly observed when fine wires or points are used as electrodes. If the point or wire is positive, the discharge generally resembles a well-defined, closely fitting, purplish sheath. It is probably a region of intense ionization produced by electrons formed in the gas and drawn to the electrode by the intense field surrounding it. With cylindrical symmetry the field is inversely proportional to the distance from the center of the wire; therefore, when the radius of the wire is small, the field in the air near it can become very large. If the

[1] LOEB, *Rev. Mod. Phys.,* **8**, 267 (1936).

small electrode is negative, the discharge is generally more reddish and it is apt to be localized at a series of points along a wire or at the extreme tip of a point. Small, branching, treelike discharges appear to grow from these points. It suggests intense, localized, positive-ion bombardment which liberates electrons from minute regions. When these are repelled from the electrode by the intense field, they produce the visible ionization in the gas. There are alternate light and dark spaces surrounding a negative point resembling in miniature the cathode glow, dark spaces, etc. The corona discharge occurs well below the sparking potential if the electrodes are very far apart in comparison with their linear dimensions. However, if this condition is not fulfilled, the inception of corona may bring with it the complete spark breakdown. To avoid corona in high-voltage work, large radii of curvature and large separations should be used. High-potential transmission lines are generally constructed with hollow tubing as this presents a much smaller external curvature than a wire with the same cross-sectional area for conduction. For long lines a central hempen core may be used to support the weight of the conductor. Corona is one of the most serious sources of power loss in high-voltage transmission and it must be carefully guarded against if the full advantages due to reduced joule heating are to be realized.

There is, however, one important application of corona in so-called *discharge-tube counters*. These are for detecting the presence of fast electrons or ions and, indirectly, light, X-rays, and gamma radiation. Consider two electrodes in the form of a fine wire and coaxial cylinder in a gas at a few centimeters pressure. At low applied voltages, where the gradients are too small to produce ionization by collisions, all the ions produced by a single ionizing event that may occur in the counter are drawn to the electrodes, and a voltage pulse of magnitude $dV = ne/C$ appears across the resistor R and total capacity C of the circuit. Since n, the number of ion pairs formed, is proportional to the energy expended by the ionizing radiation, the size of the voltage pulse is a measure of this energy. In this region the device is operating as an ionization chamber.

When the voltage is increased above a certain threshold value V_n, the Townsend discharge region is reached and secondaries appear in the immediate neighborhood of the central wire where the field dV/dr is greatest. If A ions are formed for each primary traveling to the central wire, then $dV = Ane/C$ and A is known as the *gas amplification*. As long as A is constant, the counter is operating in the proportional region and the size of the pulse is still a measure of the energy expended by the ionizing event. The shape of the voltage pulse has changed considerably, however. In general, gas amplifications up to 10^4 can be achieved before A begins to vary.

As the voltage is further increased, A is no longer the same for small pulses as for large pulses. In this region of "limited proportionality," one can still distinguish between strong and weak ionization, and the multiplication continues to be confined to a small region near the central wire.

With further increase in voltage, the discharge spreads along the full length of the tube, and all pulses are of the same height. When operating under these conditions, the counter is said to be of the Geiger-Muller type. A single electron traveling toward the central wire produces an "avalanche" of ions near the wire. The electrons are quickly drawn to

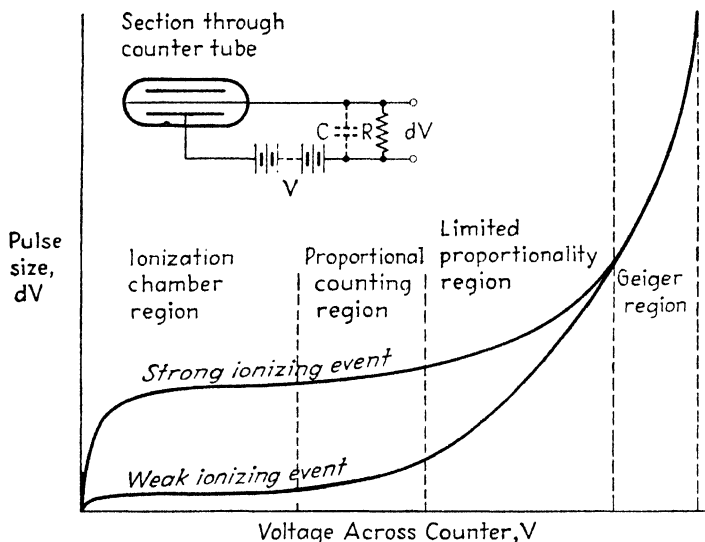



FIG. 8.16.—Typical curve illustrating the dependence of the size of voltage pulse from a counter on the voltage applied to the electrodes.

the wire, and a large positive-ion sheath is formed. When this sheath grows large enough, it reduces the field at the wire below the point where further multiplication can take place, and the avalanche stops. The positive-ion sheath will, of course, drift slowly outward until it strikes the cylinder. It is necessary that no new avalanche be initiated by absorption of photons emitted from the sheath or by secondaries emitted when the ions strike the cylinder. There are two methods of bringing this about, or "quenching" the counter, and they have led to two groups of Geiger counters called *self-quenching counters* and *non-selfquenching counters*. In the nonself-quenching case, R is large and the charge cannot leak off the wire before the sheath reaches the cylinder. The voltage is below threshold all this time. Only noble gases are used in the counter. In some cases an auxiliary circuit is employed to remove or reduce the supply voltage while the positive ions are being collected. In

the self-quenching case R is small, and as the positive sheath moves outward, the field at the wire recovers to the value necessary for discharge before the sheath has reached the cylinder. The counter contains a small trace of a polyatomic vapor (usually alcohol), which has a strong absorption for the photons emitted in the sheath and which prevents secondary emission when the positive ions strike the cylinder.[1]

The complete breakdown of a gas at high pressure resulting in a high current density, hot electrodes, etc., is known as an *arc*. It probably does not differ fundamentally from high current discharges at low pressures. The cathode-emission mechanism plays an important role and these discharges are generally characterized by a high temperature of the cathode and surrounding gas and a low-potential drop. An arc of this type can be drawn in mercury vapor at atmospheric pressure. The cathode phenomena have been already discussed in connection with mercury-vapor rectifiers. With high current densities the illumination is very intense, also its efficiency as a light source is high. As there is a relatively small surface area for the dissipation of a large amount of power, elaborate cooling precautions must be observed. A quartz envelope must be used to withstand the extreme conditions encountered.

The most familiar arcs are those that are drawn in air between solid conductors. The temperature of the cathode must be high to support a typical arc. Thus it is difficult to run an arc to a large block of copper because heat is so rapidly conducted away from the surface. Carbon, on the other hand, is a very convenient material. Its heat conductivity is low and its melting point high, the cathode surface can easily be raised to incandescence, and large current densities obtained. It is frequently used for illumination and though the majority of the light comes from the flaming vapors, the cathode itself, which is at a temperature of about 5000°C., contributes a considerable amount. Such an arc has a falling characteristic and the circuit must contain an appropriate external resistance for stability at the desired current. The discharge is localized in a glowing region between the electrodes, and the electrons, ions, hot vapors, etc., diffuse out into the surrounding neutral gas. If the distance between the electrodes is large, the discharge resembles a long sinuous flame between them. This flaming region has certain characteristics in common with the positive column of a low-pressure discharge. All the mechanisms of gaseous ionization are doubtless operative, and negative-ion attachment and recombination, together with diffusion of charged particles out of the arc, account for the disappearance of those particles that do not reach the electrodes. The cross section

[1] General reference: KORFF, "Electron and Nuclear Counters," D. Van Nostrand Company, Inc., New York, 1946.

of the glowing column of gas is doubtless determined by the rate of generation and loss of energy within it. The generation is proportional to the total current as the potential drop is approximately uniform in this region and the loss is probably roughly proportional to the external area of the arc. The column increases in cross section until the rate of loss and generation of energy are equal.

There is a uniform potential drop of the order of 30 volts per centimeter in the positive column of a carbon arc carrying a current of about 20 amp. per square centimeter. Hence, if the electrode separation is increased at constant current, the potential drop also increases. The potential drop is given as a function of the current approximately by the following relation:

$$V = a + \frac{b}{i} \tag{8.8}$$

where the constants a and b are linear functions of the electrode separation d, say, $a = \alpha + \beta d$, and $b = \gamma + \delta d$. For an arc in air between carbon electrodes about 1 cm. in diameter

$$\alpha = 39 \qquad \gamma = 11.7$$
$$\beta = 0.21 \qquad \delta = 1.05$$

where d is expressed in centimeters. These constants and Eq. (8.8) describe the electrical behavior of a carbon arc over a wide range of i and d. It was first suggested by Ayrton. Equation (8.8) represents a hyperbola and, of course, describes the behavior only of an established arc.[1]

Problems

1. Show that at a head-on elastic collision between an electron of mass m and an atom of mass M, which was previously at rest the fraction of the electron's energy that is transferred to the atom is given by

$$\frac{4Mm}{(M + m)^2}$$

Calculate the value of this fraction for a mercury atom.

2. Electrons of mass m and velocity v impinge on atoms of mass M. Assuming spherical symmetry, show that the latter gain on the average an amount of momentum in the forward direction per collision given by

$$\frac{Mm}{(M + m)}v$$

3. Show that at a recombination collision between an electron of energy eV and an atom of ionization potential V_i there is an amount of energy $\dfrac{eV}{\left(1 + \dfrac{m}{M}\right)} + eV_i$ that must be disposed of if the two are to remain together.

[1] SUITS, *Gen. Elec. Rev.*, **39**, 194 (1936).

4. If the probability that a charged particle will ionize an atom with which it collides is proportional to its velocity, at what potential would a proton produce additional ions in hydrogen if the ionization potential of hydrogen is 16 volts?

5. From the drift current through the plasma of a mercury arc it is known that the ionic and electronic densities are each about 10^{13} per cubic centimeter. If a positive-ion current of 0.041 amp. per square centimeter can be drawn from the plasma, calculate the effective positive-ion temperature.

6. A plane electrode at a negative potential in the plasma of the above problem is observed to be covered with a sheath 0.2 mm. thick. What is the potential difference between the electrode and the plasma?

7. The discharge in a thyratron can be extinguished by making the grid potential so negative that the surrounding sheath expands and completely separates the plasma regions surrounding the anode and cathode. Show how measurements of the grid potential for this cutoff, corresponding to a series of known space currents, can be used to verify the general form of the space-charge equation.

8. A discharge tube is built in such a form that there are two paths of unequal length between the electrodes. Explain why as the pressure is decreased the discharge will in general first take the shorter path, then the longer one, and in the third stage follow both.

9. The phanotron of Fig. 8.9 is used at 30°C. in a rectifier circuit in which the effective emf. is 8,000 volts. What is the minimum allowable load resistance? Calculate the efficiency of the rectifier (power in load to total power), neglecting any change in tube drop with current.

10. Assume that the grid of an FG-57 thyratron is biassed by a constant potential and that an effective alternating potential of 700 volts is applied in series with a resistance of 500 ohms in the anode circuit. Considering the characteristic (Fig. 8.10) as straight from -2 to -7 volts and neglecting the tube drop, calculate the power delivered to the load as a function of the grid potential over this range.

11. From Fig. 8.11 plot the nature of the anode-current wave for a grid potential wave (*a*) leading and (*b*) lagging the anode potential by $\pi/2$. Assume that the external resistance is sufficiently large that the tube drop can be neglected.

12. From Eq. (8.8) plot the characteristic of a carbon arc 0.5 cm. long. Determine graphically the current if the arc is in series with a 25-ohm resistance and an emf. of 100 volts.

13. A condenser of capacity C is placed across the terminals of a neon glow lamp. In series with this parallel combination is a resistance R and an applied potential V. If the lamp discharge starts at the potential V_1 and stops if the potential drops below V_2, show that the lamp will flash periodically with the period

$$\tau = RC \log_e \left(\frac{V - V_2}{V - V_1}\right)$$

The time of discharge of the condenser through the lamp, once the discharge has been initiated, is assumed to be negligible.

CHAPTER IX

ELECTROMAGNETIC EFFECTS OF STEADY CURRENTS

9.1. Ampère's Law.—The phenomena that have been discussed in the preceding chapters have all been based on Coulomb's law of force between stationary charges. This has been supplemented by the general principle of the conservation of charge and the conception of the polarization of a dielectric. The flow of electrons in metals and the motion of electrons and ions through gases or evacuated regions has also been considered, but only from the point of view of electrostatic forces. It is now necessary to introduce an entirely new type of force which is given the name "magnetic." Historically a superficial acquaintance with magnetic forces probably antedates our familiarity with electrostatic ones. For it is a force of this type that is responsible for the interaction of loadstones and iron or the orientation of a loadstone or magnet in the earth's magnetic field. The name magnet presumably comes from the extensive deposits of a magnetic oxide of iron in the ancient province of Magnesia. But there is evidence that the Chinese were familiar with the properties of the loadstone and actually used it as an aid to navigation many centuries before it was so employed in the Western Hemisphere. The first extensive scientific investigation of the properties of magnets we owe to William Gilbert who was physician-in-ordinary to Queen Elizabeth. But the study of magnetism developed quite independently of the development of electricity until Oersted, in 1819, observed more or less accidentally that a current of electricity exerts a mechanical force upon a magnet. Almost immediately thereafter Ampère observed forces of a similar nature between wires carrying electric currents, and as the result of a very beautiful series of experiments during the years 1820–1825 succeeded in arriving at a set of relations from which the law of force between wires carrying currents can be derived. Owing to the directive nature of the quantities involved, this law is somewhat more difficult to comprehend than the Coulomb law, and Ampère's discovery through his experiments and their analysis has been characterized by Maxwell as "one of the most brilliant achievements in science." For a more complete account of them than can be given here reference should be made to his "Mémoire sur la theorie mathématique des phénomènes electrodynamiques" or to Maxwell's discussion.[1]

[1] MAXWELL, "Electricity and Magnetism," 3d ed., pt. IV, Chap. II, Oxford University Press, New York, 1904.

Ampère's first experiment showed that two wires very close together traversed by equal currents but in opposite directions exert an inappreciable force on a distant current or current element. Thus the effects of oppositely directed currents at the same place and of equal magnitudes annul one another. Ampère's second experiment showed that a current-carrying conductor bent or twisted in any manner is equivalent to a series of straight conductors carrying the same current and coinciding as nearly as possible with the contorted conductor. In Ampère's third experiment a wire was bent into a portion of a circular arc and supported in a horizontal plane by a radial arm, equal in length to the radius of the arc, attached to a vertical axis about which the system was free to rotate. Electrical contacts were made near the ends of the arc by bringing up from below two mercury cups with convex menisci. When a current traversed the arc and magnets or wires carrying currents were approached in any orientation, no tendency for the system to rotate could be observed. From this Ampère concluded that the force experienced by a wire carrying a current is always normal to the wire. Ampère's fourth experiment was concerned with the reaction between similarly shaped and similarly situated wires carrying currents. The experiment is essentially the adjustment of the spacing along their common axis of three coaxial circular loops of wire, all carrying the same current, so that the forces exerted by the two outer loops on the center one exactly counteract one another. If the radii of the three loops are in the ratio $1:n:n^2$, a balance is achieved when the distance between the centers of the larger and middle loop is n times that between the centers of the middle loop and the smaller one. From this it may be shown that the force exerted by one current element on another must vary as the inverse square of the distance between them, as Coulomb found to be the case for the force between charges. Consider three elements of arc, one in each of the three circuits, that subtend equal angles at the centers of the respective circuits. These may be considered as two pairs of similarly situated elements. If the lengths of these elements are A, B, and C, respectively, and their separations are a and b, as shown in Fig. 9.1, then from symmetry the forces exerted by A and C on B must balance one another if there is no motion. These forces are known to be proportional to the current strengths (which are equal) and the lengths of the elements. Hence, if $f(d)$ represents the dependence

FIG. 9.1.—Schematic illustration of Ampère's experiment.

of the force on the distance

$$ABf(a) = BCf(b)$$

From the experiment $A/B = B/C = a/b = 1/n$; hence,

$$\frac{1}{n^2}f(a) = f(na)$$

which means that $f(a)$ must be equal to a constant divided by a^2 or the force between current-carrying wire segments must vary inversely as the square of their distance apart.

Ampère's experiments dealt with closed circuits, but these forces can be observed between currents and individual moving charges. Likewise charges in motion have been shown by Rowland[1] and Adams[2] to produce magnetic effects. Hence it is reasonable to analyze Ampère's results in terms of elementary segments of a current-carrying circuit.

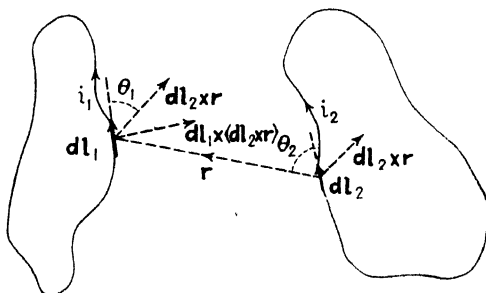



FIG. 9.2.—Illustration of the vector expression of Ampère's law.

Such an element would be written $i\,\mathbf{dl}$, where i is the magnitude of the current and $\mathbf{dl}$ is an infinitesimal vector in the direction of the circuit or current. Also, it is legitimate to write $q_l u$ for i, where q_l is the linear charge density and u the velocity with which it moves. Ampère's experiments are consistent with the statement that the magnitude of the force exerted by an element $i_2\,\mathbf{dl}_2$ on an element $i_1\,\mathbf{dl}_1$ in a vacuum (or for practical purposes in air) is proportional to each of the elements and to both $\sin\theta_1$ and $\sin\theta_2$ of Fig. 9.2 and inversely proportional to the square of their separation. Both the magnitude and direction of the force can be described most succinctly by means of the concept of the vector product which is discussed in Appendix *D*. The vector product of two vectors **A** and **B**, which is written $\mathbf{A} \times \mathbf{B}$, is a vector of magnitude $AB \sin\theta$, where θ is the angle between **A** and **B**, perpendicular to both **A** and **B** and so directed that a right-hand screw advancing in the direction

[1] ROWLAND, *Ber. d. Berl. Acad.*, p. 211 (1876).
[2] ADAMS, *Am. J. Sci.*, **12**, 155 (1901).

$\mathbf{A} \times \mathbf{B}$ would rotate the vector $\mathbf{A}$ through the smaller angle into the position occupied by $\mathbf{B}$. In terms of this concept Ampère's law of force would be written

$$\mathbf{dF}_1 = Ci_1i_2\frac{\mathbf{dl}_1 \times (\mathbf{dl}_2 \times \mathbf{r}_1)}{r^2}$$

C is an arbitrary constant which depends upon the units chosen, $\mathbf{r}_1$ is a unit vector in the direction from $\mathbf{dl}_2$ to $\mathbf{dl}_1$, and the parentheses indicate that the inner vector product is to be formed first. In the *electromagnetic system* of units (emu.), which is widely used in scientific work, C is chosen as unity. In the absolute practical system of units which is here employed the constant C is written as $\mu_0/4\pi$, where μ_0, which is known as the *permeability of free space*, has been assigned the numerical value of $4\pi \times 10^{-7}$ or 1.257×10^{-6} and the dimensions of henry per meter. Since the force is in newtons and the current in amperes, μ_0 must of necessity be in newtons per ampere², but since a newton meter is a joule or a volt-ampere second, this quotient can be written (volt-second)/(ampere-meter) or from the definition of μ_0 a henry is a volt-second per ampere. In these units Ampère's fundamental law would be written

$$\mathbf{dF}_1 = \frac{\mu_0}{4\pi}i_1i_2\frac{\mathbf{dl}_1 \times (\mathbf{dl}_2 \times \mathbf{r}_1)}{r^2} \tag{9.1}$$

or since $\mathbf{grad}\left(\frac{1}{r}\right) = -\frac{\mathbf{r}}{r^3} = -\frac{\mathbf{r}_1}{r^2}$, this equation can be written

$$\mathbf{dF}_1 = -\frac{\mu_0}{4\pi}i_1i_2\,\mathbf{dl}_1 \times \left(\mathbf{dl}_2 \times \mathbf{grad}\frac{1}{r}\right) \tag{9.2}$$

By analogy with the introduction of the electric field $\mathbf{E}$ in electrostatics it is convenient to introduce a vector $\mathbf{B}$, known as the *magnetic induction*, which determines the force on a current element. The element of induction is defined by the equation

$$\mathbf{dF}_1 = i_1\,\mathbf{dl}_1 \times \mathbf{dB}_2$$

On comparing this with Eq. (9.1) it is seen that

$$\mathbf{dB}_2 = \frac{\mu_0}{4\pi}i_2\frac{\mathbf{dl}_2 \times \mathbf{r}_1}{r^2}$$

The total induction $\mathbf{B}$ is the vector sum of all the elements $\mathbf{dB}$. If it is due to one closed circuit

$$\mathbf{B} = \frac{\mu_0}{4\pi}i\oint\frac{\mathbf{dl} \times \mathbf{r}_1}{r^2} \tag{9.3}$$

where the subscripts have been dropped and the symbol $\oint$ indicates the integration over all the elements in the closed circuit. If there are a number of current-carrying circuits, the total induction is obtained by forming the vector sum of the induction due to each circuit separately. Thus the total magnetic force on an element of a circuit or on a moving charge would be written

$$\mathbf{F}_{dl} = i\, \mathbf{dl} \times \mathbf{B}$$

or

$$\mathbf{F} = q\, \mathbf{u} \times \mathbf{B} \tag{9.4}$$

The unit of induction is evidently the newton per ampere-meter which is the same as the volt-second per square meter. The volt-second is called the *weber*, hence **B** is measured in units of weber per square meter.

9.2. Motion of Charged Particles in Magnetic and Electric Fields.—Equation (9.4) is in a very convenient form for the discussion of the motion of charged point masses such as electrons or positive ions in a region of magnetic induction. A study of the trajectories of these particles is very important, for on it depends our knowledge of electronic and ionic masses. The methods of calculating or measuring the induction will be taken up later; here it will be assumed that this quantity is known and its effect on the motion of charged particles will be considered. The results can be verified by means of a collecting electrode, connected to an electrometer or galvanometer, which can be moved around in the region traversed by the ions. Or alternatively, if there is a small amount of residual gas in the region and the ions have sufficient energy, they will excite and ionize gas molecules in their path and the radiation subsequently emitted will render the trajectories visible.

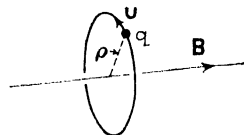



FIG. 9.3.—Description of a circular orbit of radius ρ by a positive charge moving with a velocity **u** perpendicular to the induction **B**.

From Eq. (9.4) it is evident that the force is normal to the velocity, hence **u** does not change in magnitude and the kinetic energy of the particle is constant. Consider first that **B** is a constant and that the velocity is perpendicular to **B**; in this case the force is quB normal to both **u** and **B** and since there is no motion in that direction it must be exactly counterbalanced by the centrifugal force due to the changing direction of **u**. This is mu^2/ρ, where m is the mass of the particle and ρ the radius of curvature at the point. But m, u, q, and B are all constant, so ρ is constant also and the particle describes a circular path with a radius

$$\rho = \frac{mu}{qB} \tag{9.5}$$

The angular velocity with which the path is described is equal in magnitude to u/ρ; hence it is also constant and equal to Bq/m. The axis of rotation is parallel to the induction and the sense of rotation looking in the direction of **B** is counterclockwise for a positive charge and clockwise for a negative one. Thus if the vector angular velocity $\boldsymbol{\omega}$ is defined by $\mathbf{u} = \boldsymbol{\omega} \times \mathbf{r}$,

$$\omega = -\frac{q}{m}\mathbf{B} \tag{9.5'}$$

The period of rotation in the orbit is evidently independent of the velocity. If there is a component of the velocity **u** parallel to **B**, this motion is not influenced by the presence of the magnetic induction, and the combination of the two yields a helical path of constant pitch about the axis of **B**. The pitch of the helix is evidently the ratio of the component of **u** parallel to **B** to the component perpendicular to **B**.

FIG. 9.4.—Focusing of a beam of ions by means of an axial magnetic field.

Helical Focusing.—An interesting use of these helical trajectories is in magnetic focusing devices. Consider ions in an evacuated region emerging from an opening A with a velocity **u** at an angle α with the induction **B**, as shown in Fig. 9.4. If S is a fluorescent screen distant l from A, all of these ions will be brought to a focus at a point on the screen if the time of flight over the distance l is equal to the time for one complete revolution in the helix, *i.e.*,

$$\frac{l}{u \cos \alpha} = \frac{2\pi m}{Bq}$$

The effective focal length l of the system depends on u, α, B, and q/m. If the ions have fallen through a potential difference V, their velocity is $(2qV/m)^{1/2}$; thus a knowledge of V, B, l, and α will permit the determination of the ratio q/m from which m itself can be found if q is known. This method of determining q/m is not capable of great accuracy, but as a focusing system it has numerous applications. In particular, if α is small, the cosine is approximately unity; hence for a cone of small solid angle a moderately good focus is obtained if

$$l = \frac{2\pi}{B}\left(\frac{2mV}{q}\right)^{1/2}$$

Resonance Determination of Specific Electronic Charge.—One of the most accurate methods of determining the ratio of the charge to the mass of an electron utilizes an evacuated region of constant B and a high-frequency alternating potential.[1] It is illustrated schematically in Fig. 9.5.

[1] DUNNINGTON, *Phys. Rev.*, **43**, 404 (1933).

Electrons emitted by the filament are drawn by a constant electric field to the slits S_1. A constant B deflects them in a circle, and if they have the proper energy, they enter the slits S_2 and reach the collector C. The current they constitute at this electrode is measured with an electrometer. A high-frequency potential is applied between the filament-collector assembly A' and the rest of the system A. If the electrons leave S_1 near the peak of a cycle, B is so adjusted that they will reach S_2. But if they have traversed the path in a time equal to one cycle of the potential, they will have just sufficient energy to reach A' and if a small retarding potential is applied between C and A', they cannot enter the collector C. The time of flight is θ divided by the angular velocity Be/m, and if this is equal to the period of the potential, a minimum of current will be observed to the collector C. This occurs at $\tau = m\theta/eB$, or in terms of the frequency ν

FIG. 9.5.—Precision method for determining the ratio of the charge of an electron to its mass.

$$\frac{e}{m} = \frac{\nu\theta}{B}$$

ν, θ, and B can all be measured with great accuracy and the value obtained for the charge to mass ratio is

$$\left(\frac{e}{m}\right)_{\text{electron}} = 5.274 \times 10^{17} \text{ esu./gm.}$$
$$= 1.759 \times 10^{11} \text{ coulombs/kg.}$$

Magnetic Resonance Accelerators.—The application of an alternating accelerating potential with the angular frequency Bq/m between electrodes in a region of magnetic induction B is used to produce very high-energy ions for nuclear disintegration and research. The first device employing this principle was the cyclotron.[1] The accelerating electrodes are the two halves of a shallow metal cylinder divided along a diameter as shown in Fig. 9.6. The halves are known as D's, and the entire region containing them is evacuated. The axis of the cylinder is parallel to **B** and perpendicular to the plane of the diagram. A high-frequency alter-

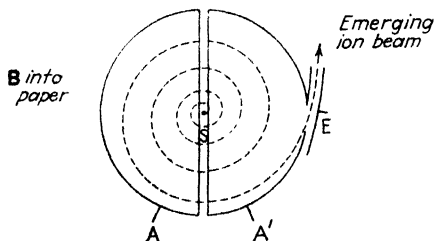



FIG. 9.6.—Schematic diagram of cyclotron ion resonance accelerator.

[1] LAWRENCE and LIVINGSTON, *Phys. Rev.*, **37**, 1707 (1931); LIVINGSTON, *Rev. Sci. Instruments*, **7**, 55 (1936).

nating potential is applied to the two halves at A and A', and ions that are produced at S are accelerated across the gap into one half or the other. Inside either electrode they are subject only to the magnetic induction which rotates them in a circle with an angular velocity Bq/m. If the time of description of a semicircle is equal to half a period of the alternating potential, they will again be accelerated on crossing the line of separation. Thus, if $\nu = Bq/2\pi m$, the ions will be continuously accelerated in an outward spiral and will achieve a final energy of $\frac{1}{2}mu^2 = (RBq)^2/2m$, where R is the final radius. This corresponds to the energy gained in falling through a potential $(RB)^2q/2m$. q is of course equal in magnitude to an integral multiple of the electronic charge e. The ions are extracted from the apparatus by a negatively charged electrode E.

An upper limit to the energy of the ions that can be produced by a cyclotron is imposed not only by R and B but also by the fact that the effective mass increases with energy in accordance with the relativistic equation $U = mc^2$. Thus as the energy increases, the effective mass increases, and by Eq. (9.5′) it is evident that the synchronous relation for constant ω and B will be violated. To overcome this difficulty either ω or B may be varied to maintain the synchronous relation for ions starting at a particular time. Such a device is known as a *synchrocyclotron* or *synchrotron*.[1] In the synchrotron B is varied, and further discussion of this instrument will be postponed till Sec. 10.6. In the synchrocyclotron the frequency is varied, generally by means of a rotating variable condenser. The feasibility of this instrument depends on the stability of the ion orbits. As the ions must complete many rotations to reach their final energy, the great majority of them would be lost by scattering with residual gas atoms or would strike the D's or walls owing to slight misalignment at starting if there were no restoring forces to maintain them in stable orbits. If a beam of ions moving in a circular path of radius given by Eq. (9.5) is considered, it may be shown that for any displacement, either radial or parallel to $\mathbf{B}$, the beam tends to return to its initial median plane and appropriate radius if the induction varies as the inverse nth power of r, where $0 < n < 1$. The ion beam is also stable for a displacement in angle about its axis of rotation or, in other words, a displacement of its phase relative to the applied alternating accelerating field between the D's. If the ions should circulate too rapidly, they would arrive at the dividing plane too early and be slowed down by the field then existing; or if they lag owing to loss of energy by collision or radiation, they traverse the dividing plane at the appropriate phase to be accelerated. Thus the ideal conditions for stable continuous acceleration

[1] MCMILLAN, *Phys. Rev.*, **68**, 143 (1945); VECKSLER, *J. Phys. U.S.S.R.*, **9**, 153 (1945); BOHM and FOLDY, *Phys. Rev.*, **70**, 249 (1946).

exist throughout one-half of the frequency-modulated cycle. Ions of heavy hydrogen (deuterons) have been accelerated to a kinetic energy of 190 million electron volts by this means.

Mass Spectrometers.—If the energy or velocity of the entering ions is known, a constant magnetic field can be used for precision determinations of the ratio q/m for positive ions. One such arrangement is shown in Fig. 9.7. Ions of velocity u enter normally an evacuated region of constant B; they are deflected through a semicircle and impinge on a photographic plate P. Wherever they strike they produce a latent image; hence a knowledge of the dimensions of the apparatus and the position of the plate enable ρ to be determined. The deflection through an angle π has a focusing property as indicated by the diagram. The central path strikes the plate the greatest distance from A. Paths making an angle δ with this one are nearer A by a factor $(1 - \cos \delta)$, and for small δ's this is small of the order δ^2. Hence the photographic trace is very narrow and has a sharp edge away from A so that its position can be measured with great precision. It is difficult to measure B accurately, so in precision measurements the instrument is used for the comparison of the ratio q/m for different ions. If B and u are known, an absolute value of q/m is obtained from Eq. (9.5), or if the potential through which the ions have fallen before entering the field is known, u may be eliminated by means of the relation $\frac{1}{2}mu^2 = qV$, yielding,

FIG. 9.7.—Schematic representation of a precision mass spectrograph for measuring q/m.

$$\frac{q}{m} = \frac{2V}{(B\rho)^2}$$

For work in atomic physics B is generally expressed in gauss (1 gauss = 10^{-4} weber per square meter), q is expressed as the number n of electronic charges, ρ in centimeters, and m in atomic-weight units on the basis $M_{\text{oxygen}} = 16$. In these units the equation becomes

$$\rho_c B_g = 144.5\left(\frac{M}{n}\right)^{\frac{1}{2}} V^{\frac{1}{2}}$$

$\rho_c B_g$ is plotted as a function of V for representative values of the parameter M/n in Fig. 9.8. From these curves the radius of curvature of the trajectory of an ion of given atomic weight in a field of B_g gauss and with an energy of V volts can be readily determined.[1]

[1] For a more complete account of the methods of measuring atomic and molecular masses the reader is referred to Aston, "Mass-spectra and Isotopes," Longmans,

Equation (9.4) gives the force on an ion in a region of magnetic induction. If there is also an electric field present, the electric force $q\mathbf{E}$ must be taken into account as well, yielding the complete force equation for a charged mass point

$$\mathbf{F} = q(\mathbf{E} + \mathbf{u} \times \mathbf{B}) \tag{9.6}$$

The integration of this equation determines the position and velocity of the particle at any time if the initial position and velocity are known. In the general case in which $\mathbf{E}$ and $\mathbf{B}$ are functions of the spatial coordinates the solution of the problem is very involved. A special instance of importance is that for which $\mathbf{E}$ and $\mathbf{B}$ are constant throughout the region in which the ion moves. The discussion is simplified by writing

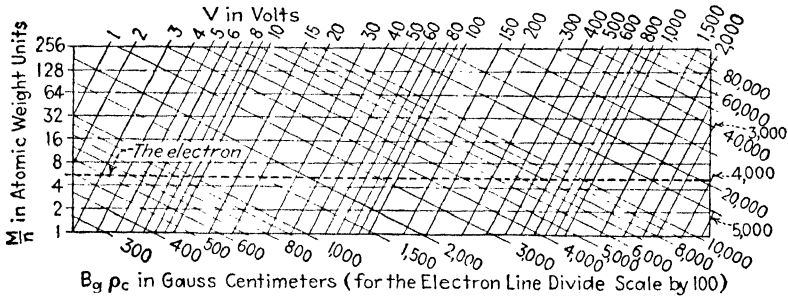



FIG. 9.8.—Logarithmic plot of the relation $B_g\rho_c = 144.5\sqrt{M/n}\sqrt{V}$.

$\mathbf{u} = \mathbf{u}' + \mathbf{U}$, where $\mathbf{U}$ is a constant velocity perpendicular to $\mathbf{E}$ and $\mathbf{B}$. Substituting this value of $\mathbf{u}$ in Eq. (9.6), it becomes

$$\mathbf{F} = q(\mathbf{E} + \mathbf{U} \times \mathbf{B} + \mathbf{u}' \times \mathbf{B})$$

Neglecting any component of $\mathbf{E}$ along $\mathbf{B}$, as this would lead merely to a uniform acceleration in the direction of $\mathbf{B}$, $\mathbf{U}$ may be chosen such that $\mathbf{E} = -\mathbf{U} \times \mathbf{B} = \mathbf{B} \times \mathbf{U}$. The force equation reduces to the form of Eq. (9.4) with $\mathbf{u}'$ occurring in place of $\mathbf{u}$. Hence the motion can be thought of as taking place in the absence of any electric field perpendicular to $\mathbf{B}$ but in a coordinate system that is moving with a velocity $\mathbf{U}$. As these three vectors, $\mathbf{U}$, $\mathbf{E}$, and $\mathbf{B}$, are mutually perpendicular, $\mathbf{U}$ is a vector in the direction $\mathbf{E} \times \mathbf{B}$ and of magnitude E/B.

The general characteristics of the motion are indicated in Fig. 9.9. From Eq. (9.5) the radius of curvature of the path in the moving system is $\rho' = mu'/qB$ and the angular velocity of rotation in the orbit is $\omega = qB/m$. The linear translation U can be written as $r'\omega$, where $r' = mE/qB^2$, and the trajectory of an ion having any initial position and

Green & Co., Inc., New York, 1933; Jordan and Young, *J. Applied Phys.*, **13**, 526 (1942); Coggeshall, *Phys. Rev.*, **70**, 270 (1946).

velocity can be thought of as the trace of a point rigidly attached, at a distance ρ' from the center, to a disk of radius r' rolling in the direction $\mathbf{U}$. Such a path is trochoidal and is a prolate cycloid, cycloid, or curtate cycloid, depending on whether ρ' is greater than, equal to, or less than r'. A number of these trajectories are of particular interest. If the ion starts from rest, the initial value of $\mathbf{u}$ is zero and hence the initial value of $\mathbf{u}'$ must be $-\mathbf{U}$. If $\mathbf{u}' = -\mathbf{U}$, the magnitudes of ρ' and r' are equal and the path is the cycloid from C to C' in Fig. 9.9. As in the case of all these trajectories, the net motion of the particle is a drift in the direction $\mathbf{U}$; the particle starting from rest never achieves a greater displacement in the direction of the electric field than $2r'$. Another interesting case is the path from A to A' in the direction $\mathbf{U}$. This corresponds to $\rho' = 0$ or $\mathbf{u}' = 0$, which implies that the velocity u of the particle must be equal to U, or E/B, in the direction $\mathbf{E} \times \mathbf{B}$. Hence

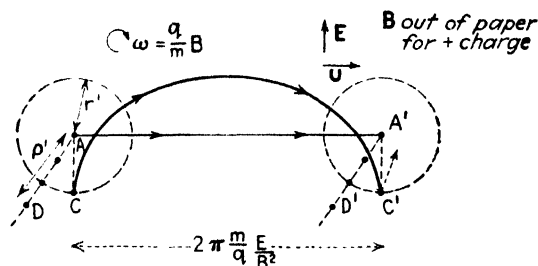



FIG. 9.9.—Motion of an ion in crossed electric and magnetic fields.

all particles regardless of mass or charge which possess this velocity move in a straight line, and if the path is defined by a series of slits, the combination of fields acts as a velocity filter or selector. A final interesting point about this system is that the imaginary disk returns to its original orientation after traversing a distance $2\pi r'$ or $2\pi mE/qB^2$. This is true regardless of the direction or magnitude of the initial velocity. Thus, if D is a source of ions, all of the ions with the proper value of q/m will be refocused upon the point D'; and the foci for different values of q/m will be distributed linearly along the line DD'. Thus the system represents a mass spectrograph, or ion selector, on the basis of q/m which is linear and has perfect focusing properties in two dimensions. In the preceding discussion only the component of $\mathbf{u}$ normal to $\mathbf{B}$ is of interest, there is no magnetic interaction between $\mathbf{B}$ and a component of $\mathbf{u}$ in its direction.

Magnetron.—A number of interesting devices depend for their action on the motion of charged particles in an evacuated region in which the magnetic induction is constant and a radial electric field $\mathbf{E}$ exists at right angles to $\mathbf{B}$. One of these devices is the simple magnetron[1] in which the

[1] HULL, *Phys. Rev.*, **18**, 31 (1921).

axis of the radial field is a long straight thermionic filament surrounded by a cylindrical plate serving as the positive electrode. These electrodes are contained in an evacuated glass envelope. If the tube is so oriented that **B** is in the direction of the filament, electrons that are drawn toward the plate are bent around by the magnetic induction; and if B is greater than a certain limiting value, they will miss the positive plate and return toward the center, resulting in no flow of current between filament and plate. This device can be used for the measurement of **B** in a region. If the simple cylindrical anode is replaced by one that has periodic resonant structures associated with it, such as the resonant-cavity magnetron, the tangential electron beam grazing the anode can excite these structures in their natural period of oscillation and very high-frequency electromagnetic radiation can be generated (Sec. 15.7). A third application of this general geometry has been used for the separation of isotopes.[1] In this apparatus the central filament is replaced by an ion beam parallel to B, which is composed of ions of the isotopes to be separated. This beam is surrounded by coaxial grids and collecting plate structures toward which the ions are accelerated by the space charge of the beam and the potentials applied to the grids and plates. The relation between maximum radius involves the mass of the ion, and hence it is possible to separate ions of different mass by suitable choice of potentials and geometry.

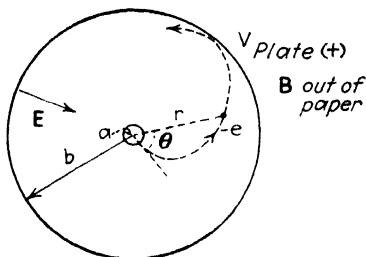



FIG. 9.10.—Section through the cylindrical plate and filament of the simple magnetron.

Figure 9.10 represents a section through the simple magnetron consisting of a thermionic filament of radius a and a collecting plate of radius b. From the discussions in connection with mass spectrometers it is evident that the motion of the ions some distance out from the center where **E** is not a rapidly varying function of r is given approximately by a circumferential velocity E/B upon which is superimposed a rotation about an instantaneous center on this circumference at an angular velocity $\omega = qB/m$. It is not, however, necessary to solve the problem explicitly for the ion orbits in order to obtain the important general relationships between the maximum radius and the other parameters. Also the following calculation is not limited to the case of a uniform radial field between a and b but applies as well in the presence of any symmetrical grid structures that do not suppress the radial-ion current. Since the magnetic field contributes no energy to the electrons, the sum of the electric potential energy $-q\int_a^r E\,dr$, where q is the ionic charge,

[1] SMITH, PARKINS, and FORRESTER, *Phys. Rev.*, **72**, 989 (1947).

and the kinetic energy is a constant which may be set equal to zero on the assumption that the ions leave the radius a from rest. Writing the energy in terms of the coordinates r and θ,

$$\frac{1}{2}m\left[\left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\theta}{dt}\right)^2\right] = q\int_a^r E\,dr$$

The torque tending to change the angular momentum of an electron about the filament is $-qBr\frac{dr}{dt}$; hence the rate of change of angular momentum is

$$m\frac{d(r^2\,d\theta/dt)}{dt} = -qBr\frac{dr}{dt}$$

Integrating

$$r\frac{d\theta}{dt} = -\frac{q}{mr}\int_a^r Br\,dr$$

Eliminating $r\,d\theta/dt$ between these equations

$$\left(\frac{dr}{dt}\right)^2 = \frac{2q}{m}\int_a^r E\,dr - \left(\frac{q}{mr}\right)^2\left(\int_a^r Br\,dr\right)^2$$

since $\mathbf{E} = -\textbf{grad}\ V$, the first integral is simply $-V_r$ if the potential of the cathode is taken as zero. The second integral reduces at once to $\frac{1}{2}B(r^2 - a^2)$ if B is constant. Thus

$$\left(\frac{dr}{dt}\right)^2 = -\frac{2qV_r}{m} - \left(\frac{qB}{2mr}\right)^2(r^2 - a^2)^2$$

If q and V_r are of opposite sign (*i.e.*, the ions or electrons would be accelerated radially outward in the absence of B), it is clear that the radial component of the velocity may vanish. Writing b for the limiting radius at which the ions are turned back ($dr/dt = 0$),

$$V_b = -\frac{q(b^2 - a^2)^2B^2}{8mb^2}$$

In the case of the simple magnetron this equation determines B at cut-off in terms of V_b, q/m for the electron, and the dimensions. In the case of the isotope separator the equation determines the mass that will reach a radius b if it is occupied by a grid at the potential V_b. Ions of greater mass will clearly achieve greater radii for the same potential distribution.

Hall and Associated Effects.—It was found by Hall in 1879 that if a current flows in a metallic strip whose plane is perpendicular to a strong magnetic field, the equipotential lines are distorted and a potential difference is developed across the strip perpendicular to the lines of flow in the absence of the field. This is accounted for qualitatively by Eq. (9.6), for the second term in that equation can be considered as

representing an induced electric field $\mathbf{u} \times \mathbf{B}$ or $\mathbf{i}_v \times \mathbf{B}/q_v$. If the electronic carriers of the current were completely free from the influence of the positive-ion lattice constituting the metal, the induced electric field should be given very closely by the previous expression with $-n_v e$ for q_v, where $-e$ is the electronic charge and n_v is the number of free conduction electrons per unit volume. This prediction is in quite good agreement with experiment for monovalent metals such as copper, silver, gold, lithium, and sodium, but even the sign is wrong for divalent metals such as beryllium, cadmium, zinc, and iron. This fact, however, is accounted for by the band theory of electrons in metals, for in accordance with this theory electrons near the top of a conduction band react to an applied force in a way that would correspond to their having a mass with a negative sign. This effectively changes the apparent sign of the charge to mass ratio and leads to a positive Hall coefficient. This coefficient for bismuth, for example, is about 10^2 times as large in magnitude as the simple free-electron theory would predict. The Hall effect is closely related to the change in conductivity of metals in a magnetic field. This effect is so large in the case of bismuth that it is actually used for the measurement of magnetic field strengths. The resistance of a bismuth wire, generally in the form of a spiral, is measured in known fields and its resistance in an unknown field can then be used to determine the field strength. There are also thermal effects associated with metallic conduction in magnetic fields. For a more detailed account of these phenomena and the theoretical attempts to account for them reference should be made to the special literature of the subject.[1]

9.3. Magnetic Field and Induction Calculations.—It is customary to use the term *field* to describe the properties of a region of space that in turn determine the physical phenomena observed to take place in the region. In Sec. 1.3 the term *electric field* was introduced and defined as the force per unit charge on an infinitesimal test charge. The term *magnetic induction* has already been employed to designate the analogous electromagnetic quantity, *i.e.*, the force experienced by a moving charge or an infinitesimal element of a current-carrying filament. In order that the concept of magnetic field may have an important extension as an auxiliary vector in describing the properties of magnetic materials, it is not defined in terms of B but rather by the equation

$$\mathbf{H} = \frac{i}{4\pi}\oint\frac{\mathbf{dl} \times \mathbf{r}_1}{\mathbf{r}^2} \tag{9.7}$$

The quantities appearing in this equation have the same significance as those in Eq. (9.3), but in the subsequent extension of the discussion to magnetic materials in Chap. XI the current i will refer specifically to the types of currents produced by emfs. with which this present discussion is alone concerned. Hence for our immediate purpose of describing electromagnetic phenomena in the absence of magnetic materials a com-

[1] CAMPBELL, "Galvanomagnetic and Thermomagnetic Effects," Longmans, Green & Co., Inc., New York, 1923; SOMMERFELD and FRANK, Thermoelectric Galvano- and Thermomagnetic Phenomena in Metals, *Rev. Mod. Phys.*, **3**, 1 (1931); SEITZ, "The Modern Theory of Solids," McGraw-Hill Book Company, Inc., New York, 1940.

parison of Eqs. (9.3) and (9.7) shows that the magnetic field and magnetic induction in free space are related in the following simple way:

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} \tag{9.8}$$

From the definition of **H** it is evident that this vector is measured in amperes per meter. It may be remarked that Eq. (9.8) holds approximately in the presence of most materials. In fact it is accurate to about 1 part in 10^5 for all substances that are commonly encountered except iron, cobalt, and nickel. The magnetic field due to a series of circuits is

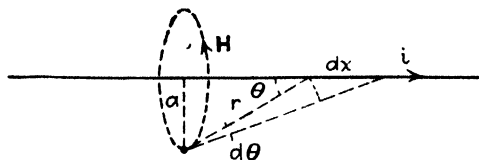



FIG. 9.11.—Magnetic field in the neighborhood of a straight current-carrying filament.

the vector sum of the **H**'s for all the circuits separately. Thus, if a volume distribution of currents is considered, the space may be thought of as filled with a great number of current filaments, and since

$$i\,\mathbf{dl} = q\mathbf{u} = q_v\mathbf{u}\,dv$$

$$\mathbf{H} = \frac{1}{4\pi}\int q_v \frac{\mathbf{u} \times \mathbf{r}_1}{r^2} dv \tag{9.9}$$

where q_v is the volume density of charge and **u** is its velocity. One of the simplest circuits that can be considered is a very long (strictly infinite) straight current-carrying filament. In this case **dl** is constant in direction and **H** must be normal to the plane defined by the filament and the point at which the field is to be calculated. From symmetry the lines of constant **H** must be circles coaxial with the wire, and from the vector-product rule the direction of **H** in one of these circles is the direction of rotation of a right-hand screw advancing in the direction of the current. For this circuit Eq. (9.7) can be written

$$H = \frac{i}{4\pi}\int_{-\infty}^{\infty} \frac{\sin\theta\,dx}{r^2}$$

From Fig. 9.11 $dx = -r\,d\theta/\sin\theta$, and a, the distance of closest approach of the filament to the point at which H is being determined, is $r \sin\theta$. Hence

$$H = \frac{i}{4\pi a}2\int_0^{\frac{\pi}{2}} \sin\theta\,d\theta$$

$$= \frac{i}{2\pi a} \tag{9.10}$$

This is known as the *Biot-Savart* law. It is a result of great practical importance, for while it is strictly true only for an infinite straight filament, it is approximately true for points sufficiently near the surface of a cylindrical wire bent into any shape, provided a, the distance to the center of the wire, is small in comparison with the linear dimensions of the complete circuit.

Another special case in which the field can be readily calculated is on the axis of a circular loop of wire, Fig. 9.12. The distance r from any element of the circuit to a fixed point on the axis is constant and equal to $(x^2 + b^2)^{1/2}$, also **r** is at right angles to **dl**. Therefore dH is given by $i\,dl/4\pi r^2$. By symmetry the component of **H** perpendicular to the axis

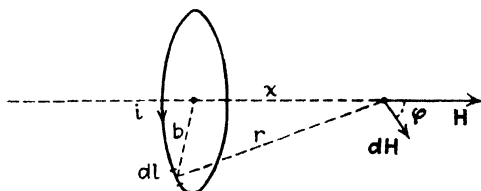



FIG. 9.12.—Magnetic field on the axis of a current-carrying loop.

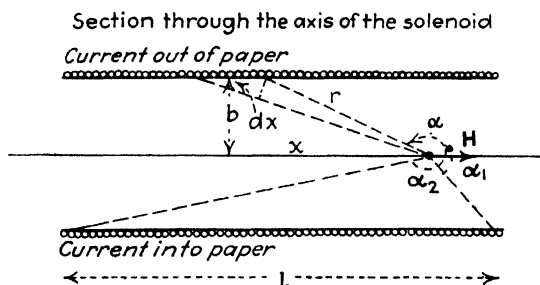



FIG. 9.13.—Calculation of the axial field of a solenoid.

vanishes for the complete circuit and the resultant **H** is in the direction of the axis and equal to $\int \cos\phi\, dH$ where $\cos\phi = b/r$. Hence

$$
\begin{aligned}
H_{\text{axial}} &= \frac{i}{4\pi}\oint \frac{b\,dl}{(b^2 + x^2)^{3/2}} \\
&= \frac{1}{2} i \frac{b^2}{(b^2 + x^2)^{3/2}} \qquad (9.11)
\end{aligned}
$$

This is a very useful expression, for from it can be calculated the axial field due to any configuration of coaxial loops. As an instance the axial field due to a solenoid, which is a closely wound helix of wire, can be found by integrating Eq. (9.11) over all the turns of the solenoid which to a first approximation can be considered as individual loops. If there are n' turns per unit length and each carries a current i, the current per length dx of the solenoid is $in'\,dx$. Therefore the axial field

at a point P is

$$H = \frac{1}{2} in' \int_l \frac{b^2\,dx}{(b^2 + x^2)^{3/2}}$$

Since $(b^2 + x^2)^{1/2}\,d\alpha = \sin\alpha\,dx$ and $b/(b^2 + x^2)^{1/2} = \sin\alpha$ in Fig. 9.13

$$\begin{aligned} H &= \tfrac{1}{2} in' \int_{\alpha_1}^{\alpha_2} \sin\alpha\,d\alpha \\ &= \tfrac{1}{2} in'(\cos\alpha_1 - \cos\alpha_2) \end{aligned}$$

At the center of the solenoid $\alpha_2 = \pi - \alpha_1$, and $H = in'l/(l^2 + 4b^2)^{1/2}$, and in the plane of one end $H = \frac{1}{2} in'l/(l^2 + b^2)^{1/2}$, where l is the length of the solenoid. At very great distances the cosines become approximately equal and the field vanishes. In the case of a solenoid which is so proportioned that l is very much greater than b, b^2 can be neglected in comparison with l^2. The central field becomes simply in' and the field in the plane of one end is just half as great. Also, if $l >> b$, the field at the center is approximately uniform over the cross section of the solenoid. While this is strictly true only for an infinite solenoid, it is approximately true for a very long straight one or for a toroidal solenoid if the radius of the toroid is much greater than the radius of its cross section.

9.4. Circuital Relations in a Magnetic Field.—There are certain properties of a magnetic field that are of great importance for the understanding of its general nature and for the calculation of fields which possess a high degree of symmetry. From Eq. (9.2) it is evident that the induction due to a closed circuit can be written

$$\mathbf{B} = \frac{\mu_0 i}{4\pi} \oint \left(\mathbf{grad}\frac{1}{r}\right) \times \mathbf{dl}$$

Let us now introduce the vector notation **curl v**, which is defined as

$$\mathbf{curl\ v} = \mathbf{i}\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right) + \mathbf{j}\left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}\right) + \mathbf{k}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)$$

The elementary properties of this vector are described in Appendix D. By forming the appropriate derivatives, or from Eq. (D.18) of the Appendix, it is seen that

$$\mathbf{grad}\,\frac{1}{r} \times \mathbf{dl} = \mathbf{curl}\left(\frac{\mathbf{dl}}{r}\right) - \frac{1}{r}\mathbf{curl\ dl}$$

However, **dl** is independent of the coordinates of the point at which **B** is measured, so the last term is zero; hence

$$B = \frac{\mu_0 i}{4\pi} \oint \mathbf{curl}\left(\frac{\mathbf{dl}}{r}\right)$$

But since the differentiation involved in the **curl** and the integration around the circuit are independent of one another, the **curl** can be taken outside the integral sign and the equation can be written

$$\mathbf{B} = \textbf{curl}\, \mathbf{A} \tag{9.12}$$

where

$$\mathbf{A} = \frac{\mu_0 i}{4\pi} \oint \frac{\mathbf{dl}}{r} \tag{9.13}$$

The fact that **B** can be written as the **curl** of a vector is a very important result. The quantity **A,** which is known as the *vector potential,* plays somewhat the same role for steady currents that the scalar potential V does in electrostatics. It is frequently somewhat easier to calculate than **B,** and once it is obtained, **B** can be found from Eq. (9.12). From the general vector relation div **curl v** = 0 (Appendix D) Eq. (9.12) can be written

$$\text{div}\, \mathbf{B} = 0 \tag{9.14}$$

The physical meaning of Eq. (9.14) is brought out by considering the integral of div **B** throughout any volume which by Stokes's theorem is equal to the integral of the normal component of **B** over the bounding surface. Equation (9.14) states that the integral of the normal component of **B** over any closed surface vanishes. Or the lines of magnetic induction, which by analogy with the lines of electric force represent by their direction and spacial density the direction and magnitude of **B,** have neither beginning nor end and may be thought of as closed curves in space.

Another valuable relation can be deduced from Eq. (9.7). Consider the result that is obtained by displacing the point at which **H** is to be calculated by an amount $\boldsymbol{\delta}\mathbf{r}$. This is equivalent to holding the point fixed and displacing each element of the circuit by an amount $-\boldsymbol{\delta}\mathbf{r}$. The scalar product of $-\boldsymbol{\delta}\mathbf{r}$ and the integrand of Eq. (9.7) can be written

$$\frac{-\boldsymbol{\delta}\mathbf{r} \cdot \mathbf{dl} \times \mathbf{r}_1}{r^2} = \frac{\mathbf{r}_1 \cdot \mathbf{dl} \times \boldsymbol{\delta}\mathbf{r}}{r^2}$$

From Fig. 9.14 $\mathbf{dl} \times \boldsymbol{\delta}\mathbf{r}$, which is equal in magnitude to $dl\, \delta r \sin\theta$, where θ is the angle between **dl** and $\boldsymbol{\delta}\mathbf{r}$, is a vector which is perpendicular to the small element of ribbon surface traced out by the displacement of the circuit. And $\mathbf{r}_1 \cdot \mathbf{dl} \times \boldsymbol{\delta}\mathbf{r}$ is the projection of this element of the surface normal to the vector **r** so the quotient of this quantity by r^2 is the element of solid angle subtended by the element of surface at the point for which **H** is being considered. The integral around the circuit is then the solid angle subtended at the point by the entire ribbon surface. Hence we may write

$$\mathbf{H} \cdot \delta\mathbf{r} = \frac{-i}{4\pi}\delta\omega \tag{9.15}$$

where $\delta\omega$, which is written for $\oint \frac{\mathbf{r}_1 \cdot d\mathbf{s}}{r^2}$, is the solid angle subtended by the ribbon surface of width δr. By Gauss's theorem, however, the integral of $d\omega$ over a closed surface (with due regard to sign) vanishes if the point from which the solid angle is considered lies outside the

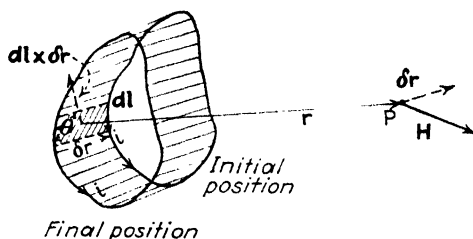



FIG. 9.14.—Expression of the scalar product of the field and displacement in terms of a solid angle.

enclosed volume. Hence, considering surfaces bounded by the two positions of the circuit and the ribbon as forming a closed drumlike surface, $\delta\omega$ must equal $-(\omega_i + \omega_f)$, where ω_i and ω_f are the solid angles subtended by the circuit in its initial and final positions, respectively.

In the previous discussions the surfaces that have been dealt with have been closed and the convention that the outward normal is positive

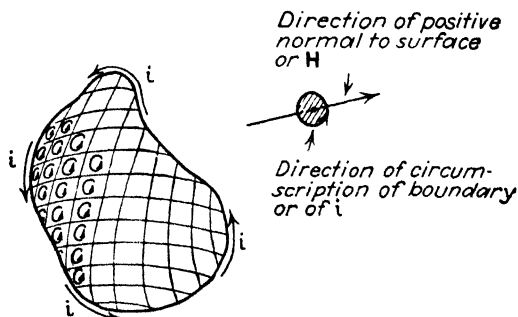



FIG. 9.15.—Subdivision of a surface bounded by a circuit into infinitesimal elements and the sign convention associated with an element.

has been adopted. In discussing the magnetic effects of electric circuits, however, surfaces are encountered which are not closed, and some other convention must be adopted to designate which of the two normals at a point shall be considered positive. It is customary to associate a sign with the normal to a surface bounded by a current-carrying circuit in the following way: Any surface bounded by a circuit can be subdivided, as indicated in Fig. 9.15, into a large number of small contiguous

elements. If a current equal to that in the bounding circuit is considered to flow around each element, the effects of coincident paths exactly cancel one another since they are traversed in opposite directions and the net effect is simply that due to the current in the bounding circuit. The positive normal to one of these infinitesimal circuits is defined as lying in the direction of advance of a right-hand screw which rotates in the sense of either the hypothetical elementary current vortex at the point or of the bounding current. As may be seen from Fig. 9.12, this is the same as the direction of the field **H** through the circuit. This convention changes the sign of ω_f in Fig. 9.14 and $\delta\omega$ can be interpreted as the difference between the solid angle subtended by the circuit in its final and initial positions. Since the solid angle subtended by the circuit is a simple scalar function of the relative position of the point and circuit

FIG. 9.16.—Potential at a great distance due to a small circuit or current vortex which is called a magnetic dipole.

positions, it is evident from Eq. (9.15) that **H** can be written as the negative gradient of the scalar function $i\omega/4\pi$, for

$$\delta\left(\frac{i\omega}{4\pi}\right) = \delta\mathbf{r} \cdot \mathbf{grad}\left(\frac{i\omega}{4\pi}\right)$$

hence on comparing this general expression with Eq. (9.15)

$$\mathbf{H} = -\mathbf{grad}\left(\frac{i\omega}{4\pi}\right) \tag{9.16}$$

The argument of the gradient is known as the *magnetic scalar potential*, Ω.

The solid angle ω subtended by the circuit is really not uniquely defined, for to it could be added any integral multiple of 4π and the same solid angle would be obtained in exact analogy to rotation about an axis. However, the gradient is independent of any such additive constant and Eq. (9.16) can be used to calculate the field if an analytical expression for the solid angle can be obtained. As a simple instance consider the field produced by a circuit at a distance which is very great in comparison with the linear dimensions of the circuit. In this case the solid angle is to a good approximation the projection of the area **a** of the circuit normal to **r** divided by r^2 or $\mathbf{a} \cdot \mathbf{r}_1/r^2$ and the potential function at a great distance is $i\mathbf{a} \cdot \mathbf{r}_1/4\pi r^2$. If a quantity known as the magnetic moment of the circuit is defined as $\mathbf{m} = \mu_0 i\mathbf{a}$, the potential may be written $\mathbf{m} \cdot \mathbf{r}_1/4\pi r^2\mu_0$. This is seen to be of the same form as the potential due to an electric dipole (Sec. 2.4) with **m** replacing the dipole moment

and μ_0 replacing κ_0. Hence such an infinitesimal current vortex is also known as a *magnetic dipole.* The lines of induction due to the circuit are the same in these distant regions as the lines of force from a dipole. This is a very useful result and will be employed in connection with various subsequent discussions.

A consideration of the integral of Eq. (9.16) around a closed path yields another important result. The integral of $\mathbf{H} \cdot \mathbf{dl}$ around any closed path is known as the *magnetomotive force* and written as $\mathcal{H}$. This is analogous to the definition of the electromotive force $\mathcal{E}$ as the integral around a closed path of $\mathbf{E} \cdot \mathbf{dl}$. The integral of $\mathbf{dl} \cdot \mathbf{grad}\ \omega$ around a simple closed path is either $\pm 4\pi$ or zero, depending on whether the path links the circuit or not. If the path does not link the circuit, the integrand runs through a series of equivalent positive and negative values and the result is zero. If the path links the circuit in the sense indicated in Fig. 9.17, the solid angle runs through a continuously increasing series of values from ω to $4\pi + \omega$ and the integral is equal to i. If the path forms n net linkages of the circuit in this direction, the result of the integration is ni. In general, the integral may be written

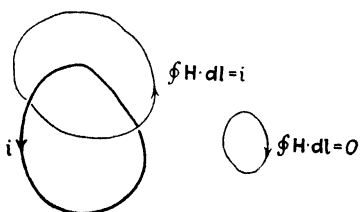



FIG. 9.17.—Circuital relation for determining the magnetomotive force.

$$\oint \mathbf{H} \cdot \mathbf{dl} = \mathcal{H} = i \tag{9.17}$$

where i is the total current that has been linked by the path of integration. This is a very important relation and may be used immediately for the calculation of the field, $\mathbf{H}$, in the cases of very symmetrical circuits. For example, the field due to a straight wire is obtained by integrating $\mathbf{H} \cdot \mathbf{dl}$ around a coaxial circle say of radius a. Since, by symmetry, $\mathbf{H}$ is constant in this path, $2\pi a H = i$ or $H = i/2\pi a$, in agreement with Eq. (9.10). If space is considered to be filled with current filaments which constitute a current density $\mathbf{i}_v$ per unit volume, Eq. (9.17) would be written

$$\oint \mathbf{H} \cdot \mathbf{dl} = \int_s \mathbf{i}_v \cdot \mathbf{ds}$$

where $\mathbf{i}_v \cdot \mathbf{ds}$ is the normal component of the current flow through an element ds of the surface bounded by the path of integration $\oint \mathbf{H} \cdot \mathbf{dl}$. From Stokes's theorem [Eq. (*D*.15) of Appendix *D*]

$$\oint \mathbf{H} \cdot \mathbf{dl} = \int_s \mathbf{curl}\ \mathbf{H} \cdot \mathbf{ds}$$

so

$$\int_s \mathbf{curl}\ \mathbf{H} \cdot \mathbf{ds} = \int_s \mathbf{i}_v \cdot \mathbf{ds}$$

As this equation is true at every point, the integrands may be equated, yielding

$$\mathbf{curl\ H} = \mathbf{i}_v \tag{9.18}$$

Equations (9.14) and (9.18) constitute the two fundamental differential equations which determine the magnetic field associated with steady currents. Equation (9.18) will require modification for changing currents, but this discussion will be postponed until the following chapter.

The magnetic scalar potential is a solution of Laplace's equation in regions where there is no current. This can be shown by considering the solid angle subtended at a point by any element of area. This is equal to $(a \cos \theta)/r^2$, where θ is the angle between the normal to the area a and the radius vector r to the point in question. This expression may be shown by substitution to be a solution of Laplace's question, and since the sum of any number of terms of this form is also a solution, Ω is a solution, *i.e.*,

$$\nabla^2\Omega = 0$$

Ω for a circular coil can be found by obtaining a solution of Laplace's equation that reduces to Eq. (9.11) on the axis.[1] The magnetic vector potential in free space is a solution of a vector equation that is analogous to the scalar equation of Poisson in electrostatics. Since $\mathbf{H} = \mathbf{B}/\mu_0 = (1/\mu_0)\ \mathbf{curl\ A}$ and (Appendix *D*) $\mathbf{curl\ curl\ A} = \mathbf{grad}$ div $\mathbf{A} - \nabla^2\mathbf{A}$, (Eq. 9.18) yields

$$\nabla^2\mathbf{A} - \mathbf{grad}\ \text{div}\ \mathbf{A} = -\mu_0\mathbf{i}_v$$

It may be shown that the divergence of **A** vanishes. In the case of a spatial distribution of currents $i\ \mathbf{dl}$ would be written $\mathbf{i}_v\ dv$ and Eq. (9.13) would become

$$\mathbf{A} = \frac{\mu_0}{4\pi}\int \frac{\mathbf{i}_v}{r}\, dv$$

the integration becoming a volume integration. Div **A** is the integral of div $(\mathbf{i}_v/r)$ times the numerical constant. $\mathbf{i}_v$, however, is independent of the coordinates with respect to which the divergence is taken; hence the integrand becomes $\mathbf{i}_v \cdot \mathbf{grad}\ (1/r)$. Letting x, y, and z be the coordinates of the field point for which **A** is to be determined and which are involved in the gradient and letting x', y', and z' specify the position of the volume element dv and current $\mathbf{i}_v$, r would be written

$$r = [(x - x')^2 + (y - y')^2 + (z - z')^2]^{1/2}$$

From this form it is evident that a partial derivative with respect to an unprimed variable is equal in magnitude but opposite in sign to the analogous derivative with respect to a primed variable. Thus

$$\text{div}\ \mathbf{A} = \frac{-\mu_0}{4\pi}\int \mathbf{i}_v \cdot \mathbf{grad}'\, \frac{1}{r}\, dv$$

On rewriting the integrand by means of the identity

$$\text{div}'\, \frac{\mathbf{i}_v}{r} = \frac{1}{r}\, \text{div}'\, \mathbf{i}_v + \mathbf{i}_v \cdot \mathbf{grad}'\, \frac{1}{r}$$

and transforming the volume integral of the divergence to the surface integral of

[1] For the details of this method see, for instance, Maxwell, *loc. cit.*

$\mathbf{i}_v/r$, the equation becomes

$$\operatorname{div} \mathbf{A} = \frac{-\mu_0}{4\pi}\left(\int_s \frac{\mathbf{i}_v}{r} \cdot d\mathbf{s} + \int_v \frac{1}{r} \operatorname{div}' \mathbf{i}_v \, dv\right)$$

Since div $\mathbf{i}_v = 0$, the second term vanishes and the bounding surface may be chosen at a great distance so as to be beyond all currents and the first term vanishes. Thus div $\mathbf{A} = 0$ and the original differential equation can be written

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{i}_v$$

This vector equation represents three scalar equations in the components and is the analogue of Poisson's equation in electrostatics. At points where $\mathbf{i}_v = 0$ it becomes Laplace's equation, and $\mathbf{A}$ at a point where there is no current is a solution of Laplace's equation.

9.5. Energy Relations and Forces between Circuits.—In the discussion of forces between circuits, as in the discussion of forces between charged conductors, the simplest method of approach is the construction of an energy function. In this discussion it will be assumed that the magnetic induction $\mathbf{B}$ and the currents are maintained constant by appropriate batteries. Consider a circuit that is given a vector displacement $\delta\mathbf{r}$. By Eq. (9.4) the work done by the mechanical force on the element $d\mathbf{l}$ is

FIG. 9.18.—Calculation of the work done in displacing a current-carrying circuit in a magnetic field.

$$dW = -\mathbf{F}_{dl} \cdot \delta\mathbf{r} = -i \, d\mathbf{l} \times \mathbf{B} \cdot \delta\mathbf{r} = i \, d\mathbf{l} \times \delta\mathbf{r} \cdot \mathbf{B}$$

But from Fig. 9.18 this is seen to be the inward normal flux of $\mathbf{B}$ through the element of ribbon surface traced out by $d\mathbf{l}$ in the displacement $\delta\mathbf{r}$ into the page. The integral of this quantity around the complete circuit yields the work δW performed. Since div $\mathbf{B} = 0$, this is equal to the outward flux of $\mathbf{B}$ through the surfaces bounded by the initial and final positions of the circuit. With the right-hand-screw sign convention of the previous section the work done, δW, is given by

$$\delta W = -i\left(\int_j \mathbf{B} \cdot d\mathbf{s} - \int_i \mathbf{B} \cdot d\mathbf{s}\right)$$

where the subscripts i and j refer to the initial and final positions of the circuit. The integral of $\mathbf{B} \cdot d\mathbf{s}$ over a surface is said to be the *magnetic flux*, or simply *flux*, through the circuit. It is evidently a simple scalar quantity

$$\phi = \int \mathbf{B} \cdot d\mathbf{s} \tag{9.19}$$

In terms of the flux

$$\delta W = -i \, \delta\phi$$

Choosing $W = 0$ when $\phi = 0$, this equation may be integrated to yield $W = -i\phi$ since i is assumed to be constant. By an extension of the above argument it may be seen that the force or torque on the circuit is evidently the negative partial derivative of this energy W with respect to the appropriate coordinate. However, a magnetic energy function will be defined as $U = -W$. The circuit tends to move in such a sense as to increase U and the forces and torques are obtained by taking the positive partial derivatives with respect to U.*

$$U = i\phi = i\int \mathbf{B} \cdot \mathbf{ds} \tag{9.20}$$

A useful alternative form can be obtained from Eq. (9.12) by means of Stokes's theorem

$$\int \mathbf{B} \cdot \mathbf{ds} = \int \text{curl } \mathbf{A} \cdot \mathbf{ds} = \oint \mathbf{A} \cdot \mathbf{dl}$$

Hence,

$$U = \oint i\mathbf{A} \cdot \mathbf{dl} \tag{9.21}$$

These various forms for the energy function will be used in the subsequent discussion. It is evident from any of these expressions that a circuit carrying a constant current tends to take up such a position as to include the largest possible flux through it in the positive sense of the right-hand-screw notation.

The energy of a small constant-current vortex of moment $\mathbf{m} = \mu_0 i\mathbf{a}$ in a field $\mathbf{H} = \mathbf{B}/\mu_0$ is evidently

$$U = i\varphi = i\mathbf{B} \cdot \mathbf{a} = \mathbf{m} \cdot \mathbf{H}$$

The torque on such a vortex is given with its proper sign as

$$\mathbf{T} = \mathbf{m} \times \mathbf{H}$$

and the force on it in a field that is a function of the coordinates is

$$\mathbf{F} = (\mathbf{m} \cdot \text{grad})\mathbf{H}$$

in close analogy with the electric-dipole case of Sec. 2.4.

In dealing with currents in wires the primary interest attaches to the forces and torques on the circuits in terms of the currents they carry and the geometrical variables specifying their positions. The mutual

* This is analogous to the electrostatic energy discussion in Sec. 2.1. There the batteries that maintained the potentials of the conductors constant delivered twice the change in electrostatic energy with opposite sign that would have occurred had the charges been kept constant instead. The motion of the conductors here is assumed to take place at constant current. It will be seen in Sec. 10.1 that in such a case the batteries which maintain these currents constant deliver twice the mechanical work W with opposite sign; hence the total energy function is U instead of W.

energy between two circuits can be obtained by inserting the value of **A** from Eq. (9.13) in Eq. (9.21) and attaching subscripts to distinguish the two paths of integration

$$U_{12} = \frac{\mu_0}{4\pi} i_1 i_2 \oint_1 \oint_2 \frac{\mathbf{dl}_1 \cdot \mathbf{dl}_2}{r_{12}}$$

The integral is purely a geometrical quantity that can be calculated when the positions of the circuits are specified though the computation is frequently difficult to perform. The coefficient of $i_1 i_2$ is known as the *coefficient of mutual inductance* between the circuits and is generally written

$$L_{12} = \frac{\mu_0}{4\pi} \oint_1 \oint_2 \frac{\mathbf{dl}_1 \cdot \mathbf{dl}_2}{r_{12}} \tag{9.22}$$

This is known as *Neuman's formula.* It is evident from the symmetry of this expression that $L_{12} = L_{21}$ and in terms of this coefficient the mutual energy is written

$$U_{12} = L_{12} i_1 i_2 \tag{9.23}$$

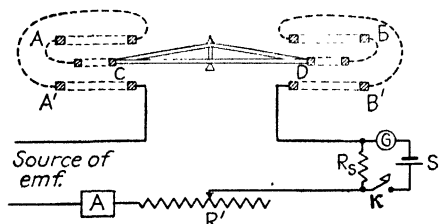



FIG. 9.19.—Schematic representation of a current balance.

The inductance is evidently measured in henries if lengths are measured in meters.

The absolute determination of the ampere in terms of the fundamental units of length, mass, and time depends on Eq. (9.23) and the calculation of a coefficient of mutual inductance by means of Eq. (9.22). If x is the separation between the centers of two coaxial coils carrying the same current, the axial force between them is given by

$$F_x = \frac{\partial U_{12}}{\partial x} = i^2 \frac{\partial L_{12}}{\partial x}$$

If this force is balanced by the gravitational force on a mass M,

$$i^2 \frac{\partial L_{12}}{\partial x} = Mg$$

hence i can be determined in terms of the geometrical factor, the mass M, and the acceleration of gravity. The latter quantity must be accurately determined at the point in question; its approximate value is 9.8 m. per second per second. The actual determinations are performed in national standardizing laboratories by means of an instrument known as a dynamometer or current balance.[1] The general principles of

[1] VIGOUREUX and WEBB, "Principles of Electric and Magnetic Measurements," Prentice-Hall, Inc., New York, 1936; CURTIS, "Electrical Measurements," McGraw-Hill Book Company, Inc., New York, 1937.

construction are indicated in Fig. 9.19. Two identical coils C and D in a horizontal plane are attached rigidly to the extremities of a balance arm. Symmetrically placed above and below each movable coil are two fixed coils which carry the same current. The coils are connected in such a way that they both tend to rotate the balance arm in the same sense. This tendency is counteracted by weights added to the proper arm a known distance from the fulcrum. The balancing procedure is the same as in ordinary weighing and the current at balance is determined in terms of the gravitational torque and the rate of change of mutual inductance between a movable coil and its neighboring stationary ones. This latter quantity can be calculated from the dimensions. The effect of stray constant fields can be eliminated by making a second measurement reversing the sense of flow of the current. The effect on one arm of the field produced by the coils surrounding the other can be eliminated by making another measurement in which the sense of flow of the current is reversed in just one group of coils, say, B, D, and B'. Of course, no magnetic materials can be used in connection with the instrument, for in that case the mutual inductance could not be calculated simply from the coil dimensions. An ammeter A can be calibrated in this way by varying the current through the balance by means of the adjustable resistance R'. The accuracy obtainable is of the order of 1 part in 10^5. In practice, standardizing laboratories generally use the balance to establish the calibration of secondary standards of potential since these are more permanent and reliable than moving-coil meters. R_s is a standard of resistance that can be established independently by a method described in Sec. 10.4. The current is varied by means of R' until the potential drop across R_s is equal to the emf. $\mathcal{E}_s$ of the standard cell. In this condition there is no deflection of the galvanometer on closing the key K. The value of i is determined from the mass on the balance and $\mathcal{E}_s = R_s i$, where both R_s and i are known. The standard cells are then used as standards of potential in ordinary laboratories. These may be used to calibrate voltmeters and, in conjunction with standards of resistance, to calibrate ammeters, which is essentially the reverse of the procedure outlined above.

Equation (9.23) represents the mutual magnetic energy between any two circuits. In the general case of n circuits, this expression can evidently be generalized to

$$U = \frac{1}{2} \sum_{j=1}^{j=n} \sum_{k=1}^{k=n} L_{jk} i_j i_k \tag{9.24}$$

where the double summation over the circuits excludes the case of $j = k$ for this is the self-energy of a particular circuit which has not yet been considered. The factor $\frac{1}{2}$ occurs because in the double summa-

tion each circuit is counted twice, and since $L_{jk} = L_{kj}$, pairs of corresponding terms in Eq. (9.24) reduce to the form of Eq. (9.23). The energy of a single circuit carrying a current can be written down formally from Eq. (9.20) or (9.21). Since B or A is proportional to the current, the energy U associated with a current i flowing in a wire is proportional to i^2. In order to conform to the terms of Eq. (9.24) the constant of proportionality is written $\frac{1}{2}L$. Thus

$$U = \tfrac{1}{2}Li^2 \tag{9.25}$$

The constant L is known as the *coefficient of self-inductance.* If there are a number of circuits, the self-inductance of circuit i is written L_{ii}. Methods of calculating L or L_{ii} will be discussed in Sec. 9.6. The total magnetic energy represented by the n current-carrying circuits is given by Eq. (9.24) plus the self-energies of the form of Eq. (9.25) for each of the individual circuits. This total sum can evidently still be written as Eq. (9.24) but without excluding the terms of the form $\frac{1}{2}L_{jj}i_j^2$, for these are exactly the self-energies of the separate circuits. Hence this expression represents the total energy in the magnetic form with the understanding that all the terms are included. Any component of force or torque can be obtained by taking the positive partial derivative of U with respect to the appropriate variable. As an instance in the case of a single circular circuit the force tending to increase its radius is

$$F_r = \frac{\partial U}{\partial r} = \frac{1}{2}i^2\frac{\partial L}{\partial r}$$

The energy associated with a spatial distribution of currents can be obtained by the obvious generalization of Eq. (9.24). The total flux through circuit j is

$$\phi_j = \sum_{k=1}^{k=n} L_{jk}i_k$$

and in terms of this flux Eq. (9.24) can be written

$$U = \frac{1}{2}\sum_{j=1}^{j=n} \phi_j i_j$$

This differs from Eq. (9.20) by the factor $\frac{1}{2}$ because that equation referred to the energy of a circuit in a previously established field. By Stokes's theorem

$$\phi_j = \int_j \mathbf{B} \cdot d\mathbf{s} = \oint_j \mathbf{A} \cdot d\mathbf{l},$$

or

$$U = \frac{1}{2}\sum_{j=1}^{j=n} \oint_j i_j \mathbf{A} \cdot d\mathbf{l}$$

This is the energy from the point of view of closed filamentary circuits that may be considered to be distributed in any way throughout a conductor or series of conductors. The sum of the current elements $i_j\,\mathbf{dl}$ in a small region may be considered to constitute the current density $\mathbf{i}_v$ times the volume element dv comprising the region. The summation and circuital integration may then be considered as a volume integration and the energy written

$$U = \tfrac{1}{2}\int_v \mathbf{i}_v \cdot \mathbf{A}\,dv \tag{9.26}$$

For greater convenience in certain types of calculation this can be written in terms of the magnetic field and magnetic induction. From Eq. (9.18)

$$U = \tfrac{1}{2}\int \mathbf{A} \cdot \mathbf{curl\ H}\,dv$$

Since $\text{div}\,(\mathbf{A} \times \mathbf{H}) = \mathbf{H} \cdot \mathbf{curl\ A} - \mathbf{A} \cdot \mathbf{curl\ H}$, the integrand can be written

$$\mathbf{H} \cdot \mathbf{curl\ A} - \text{div}\,(\mathbf{A} \times \mathbf{H})$$

But the integral of the divergence term is zero if the currents occupy only a finite region. This can be seen by transforming the integral by the theorem of flux into a surface integral of the normal component of $\mathbf{A} \times \mathbf{H}$ over a surface at a great distance. $\mathbf{A}$ is inversely proportional to r and $\mathbf{H}$ is inversely proportional to r^2, whereas the surface area increases only as r^2; so the integral vanishes as $1/r$. Writing $\mathbf{B}$ for $\mathbf{curl\ A}$ the energy can be expressed as

$$U = \tfrac{1}{2}\int \mathbf{B} \cdot \mathbf{H}\,dv$$

Or since $\mathbf{B}$ differs from $\mathbf{H}$ in the cases under consideration at present only by the constant μ_0

$$U = \frac{1}{2}\int \mathbf{B} \cdot \mathbf{H}\,dv = \frac{1}{2\mu_0}\int B^2\,dv = \frac{\mu_0}{2}\int H^2\,dv \tag{9.27}$$

Thus the total magnetic energy associated with any current configuration is equal to $1/2\mu_0$ times the integral of the square of the resultant induction throughout space. This form of the energy expression is useful in many instances. It facilitates the calculation of inductance coefficients and can be used in determining the distribution of high-frequency currents in conductors. On the general dynamical principle that the stable configuration of a group of current-carrying circuits that are free to change their positions is that for which U has the smallest value, Eq. (9.27) can be used in many instances to determine the direction and order of magnitude of a force or torque.

The electrodynamometer will be considered as an application of the ideas presented in this section. This instrument consists of two coils, one of which is free to rotate in the field of the other. The torque on the movable coil is measured by the angular displacement of a pointer against the restoring torque of a spiral spring. If this restoring torque is proportional to the displacement, the displacement is proportional to the product of the currents in the two coils and hence can be used to measure various electrical quantities. Figure 9.20 is a schematic diagram of such an instrument. From Eq. (9.23) the torque on the inner

coil is

$$T = \frac{\partial U_{ab}}{\partial \theta} = i_a i_b \frac{\partial L_{ab}}{\partial \theta} = C\theta$$

where C is the restoring torque of the spring per unit angular displacement. L_{ab} can be calculated or determined experimentally at a series of angles by sending a known current through the coils in series. In one form of the instrument b is in the form of a long solenoid. The field is then approximately constant and equal to $n'i_b$, where n' is the number of turns per meter of the solenoid. The flux through coil a times the number of turns is $\mu_0 n' i_b A n_a \sin\theta$, where A is the area of this coil. Hence

$$U_{ab} = \mu_0 i_a i_b n_a n' A \sin\theta$$

and,

$$T = \mu_0 n_a n' A \cos\theta i_a i_b$$

In any case the displacement θ is a measure of the product $i_a i_b$. If the coils are arranged in series, the deflection is a measure of i^2. Thus the angular displacement is of the same sign for either direction of i. This means that the instrument can be used for measuring alternating currents. If $i = i_0 \sin \omega t$, the average value of i^2 which is recorded by the instrument is $i_0^2/2$ or i_e^2 the square of the effective value. If the wave is a complex one, the deflection is proportional to one-half the sum of the squares of the Fourier coefficients. As shown in Fig. 9.20, the instrument is arranged as a wattmeter. R is a resistance large in comparison with that presented by the coil a. The current through a is then proportional in both magnitude and phase to the potential difference across the load. The current i_b is equal or, if a shunt is used, proportional to the current through the load; therefore the deflection of the instrument is proportional to the product Vi, which is equal to the power consumed by the load. In the case of an alternating current the instrument measures the average value of iV which has been seen to be $i_e V_e \cos\alpha$ where α is the phase difference between i and V.

FIG. 9.20.—Electrodynamometer arranged as a wattmeter.

9.6. Calculation of Coefficients of Inductance.—Coefficients of inductance are of so much importance in all electromagnetic problems that the method of calculating them will be considered in detail for a number of simple instances. Frequently it is easier to use the definition of the inductance as the flux linkage per unit current than to proceed directly from Eq. (9.22). Consider first a solenoid that is so long in comparison with its radius that the field near the center is to a good approximation uniform and equal to $n'i$. The flux through the central region is then

$\mu_0 An'i$, where A is the cross-sectional area of the solenoid. The linkage with a coil of m turns wound around the solenoid near its center is then $\mu_0 An'mi$, or the mutual inductance between them is

$$L_{mn} = \mu_0 An'm \tag{9.28}$$

where A is the area of the solenoid in square meters and n' is the number of turns on it per meter.

Another instance which can be handled in this way is the circular solenoid or toroid. Figure 9.21 represents a section through the axis of a toroid of radius b and cross-sectional radius a. By symmetry the lines of induction are circles coaxial with the toroid. Consider the magnetomotive force in one of these circles of radius $b - r\cos\theta$.

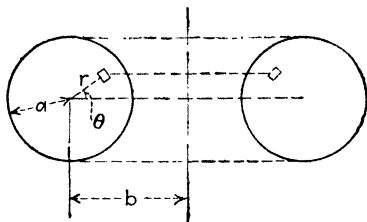



FIG. 9.21.—Section through a toroid for the calculation of the coefficients of inductance.

$$\mathcal{H} = 2\pi(b - r\cos\theta)H = ni$$

where n is the number of turns on the toroid and i is the current through the wire. Since $B = \mu_0 H$

$$\phi = \frac{\mu_0 ni}{2\pi} 2\int_0^a \int_0^\pi \frac{r\,dr\,d\theta}{(b - r\cos\theta)}$$

The θ integral is found from tables to be $\frac{\pi}{(b^2 - r^2)^{1/2}}$; hence

$$\phi = \mu_0 ni \int_0^a \frac{r\,dr}{(b^2 - r^2)^{1/2}} = \mu_0 ni[b - (b^2 - a^2)^{1/2}]$$

The mutual inductance between such a toroid and a coil of m turns wound upon it is then

$$L_{mn} = \mu_0 mn[b - (b^2 - a^2)^{1/2}]$$

The self-inductance of the toroid itself is

$$L = \mu_0 n^2[b - (b^2 - a^2)^{1/2}]$$

If $b >> a$, the bracket can be expanded neglecting $(a/b)^4$ and the self-inductance written

$$L = \frac{1}{2}\mu_0 n^2 \frac{A}{\pi b} = \mu_0 n^2 \frac{A}{l} \tag{9.29}$$

where A is the cross-sectional area of the toroid and l is the length of the central circle.

The self-inductance of the two common types of transmission lines can also be calculated by this method. In the case of two small parallel

wires a distance d apart carrying equal currents in opposite directions, the lines of induction are the same as the equipotentials in the electrostatic problem of two oppositely charge infinite cylinders. This can be shown from the Biot-Savart law. A cross section through this system

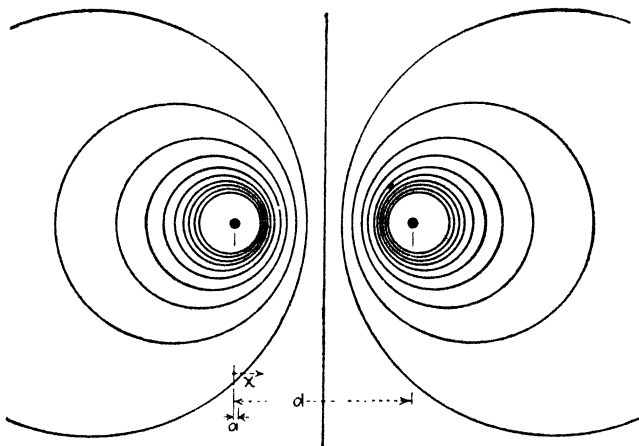



FIG. 9.22.—Lines of induction around parallel wires carrying equal currents in opposite directions (innermost circles are not drawn in).

perpendicular to the wires is shown in Fig. 9.22. If the radius of the wires, a, is much less than d, the flux in the wires themselves can be neglected. The flux per meter length due to both wires passing normally through a plane surface extending from one wire to the other can be written

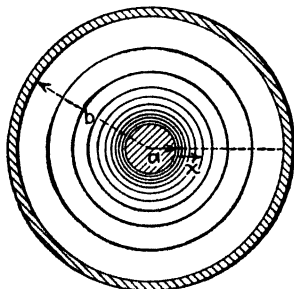



FIG. 9.23.—Calculation of the inductance per unit length of a coaxial line.

$$\phi = 2\frac{\mu_0 i}{2\pi}\int_a^d \frac{dx}{x} = \frac{\mu_0 i}{\pi}\log_e\frac{d}{a}$$

Or the self-inductance of the system per unit length is

$$L' = \frac{\mu_0}{\pi}\log_e\frac{d}{a} \tag{9.30}$$

Another important case is the so-called *coaxial line.* This consists of a pair of coaxial cylindrical conductors carrying equal currents in opposite directions. Since the currents are equal and opposite, the value of $\mathcal{H}$ around a coaxial circle external to both conductors must be zero. From the symmetry of the system this implies that $\mathbf{H}$ is everywhere zero in this region. If the radius of the inner conductor, a, is small in comparison with that of the outer conductor, b, only the flux in the region between the two need be considered. From the Biot-Savart law,

$$\varphi = \frac{i}{2}\frac{\mu_0}{\pi}\int_a^b \frac{dx}{x} = \frac{\mu_0 i}{2\pi}\log_e\frac{b}{a}$$

hence the self-inductance per unit length of this system is

$$L' = \frac{\mu_0}{2\pi}\log_e\frac{b}{a} \tag{9.31}$$

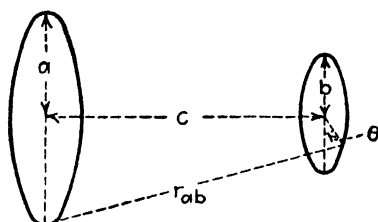
For most other inductance calculations it is necessary to resort to Neuman's formula [Eq. (9.22)]. For a discussion of the various calculable cases reference should be made to treatises dealing with this subject.[1] As one instance of the method consider the mutual inductance between two coaxial circles of radii a and b with their centers a distance c apart. The element of length of circuit b can be written $dl_b = b\,d\theta$ and the scalar product of the two as $b \cos\theta\, dl_a\, d\theta$. The distance from dl_a to an element on circuit b is seen from the geometry to be

$$r_{ab} = (c^2 + a^2 + b^2 - 2ab\cos\theta)^{1/2}$$

FIG. 9.24.—Calculation of the mutual inductance of two coaxial circles.

Since all elements of a are alike, the integral of dl_a reduces simply to $2\pi a$. Therefore Eq. (9.22) becomes

$$L_{ab} = \tfrac{1}{2}\mu_0 ab\int_0^{2\pi}\cos\theta(c^2 + a^2 + b^2 - 2ab\cos\theta)^{-1/2}d\theta$$

This is an elliptic integral and cannot be expressed in terms of simple functions. However, it can be written

$$L_{ab} = \mu_0(ab)^{1/2}\left[\left(\frac{2}{k} - k\right)K - \frac{2}{k}E\right]$$

where $k^2 = 4ab[(a+b)^2 + c^2]^{-1}$ and K and E are the elliptic integrals

$$K = \int_0^{\frac{\pi}{2}}\frac{d\phi}{(1 - k^2\sin^2\phi)^{1/2}} \qquad E = \int_0^{\frac{\pi}{2}}(1 - k^2\sin^2\phi)^{1/2}\,d\phi$$

which can be found from tables. There are two special cases that can be expressed more simply. If $c >> a$ and b.

$$L_{ab} = \frac{\mu_0\pi}{16}(ab)^{1/2}k^3$$

[1] GRAY, "Absolute Measurements in Electricity and Magnetism," The Macmillan Company, New York, 1921.

and if $c << a$ and b, and a is approximately equal to b

$$L_{ab} = \mu_0 a\left(\log_e \frac{8a}{d} - 2\right)$$

where $d = [(a - b)^2 + c^2]^{1/2}$ is the distance of closest approach of the loops. If the coils are of m and n turns, respectively, these values of the mutual inductance must be multiplied by the product mn.

The self-inductance of a single circular loop of wire of radius R and cross-sectional radius r can be obtained from this result. For the purpose of calculating the magnetic energy from Eq. (9.27) consider that the integration is carried out in two parts: (1) the region external to the wire, (2) the region of the wire itself. Region 1 is assumed to be very much greater than region 2. In region 1 it will be legitimate to consider all the current as concentrated in the central filament of the wire of radius of curvature R. From the previous result the energy associated with the current i and the flux through region 1 must be equal to $\frac{1}{2}L_1 i^2$, where L_1 is the mutual inductance between this central filament and any filament lying in the external boundary of the wire. The energy in the wire may be calculated as follows: The current density (uniformity is assumed) is $i/\pi r^2$. $\mathcal{H}$ around any circular path inside the wire coaxial with its central filament and a distance x from it is

$$\mathcal{H} = 2\pi x H = i\frac{x^2}{r^2}$$

Thus $H = ix/2\pi r^2$ and the integral of H^2 throughout the volume of the wire is

$$\int H^2\, dv = \frac{i^2}{4\pi^2 r^4} 2\pi R \int_0^r 2\pi x^3 dx = \frac{i^2 R}{4}$$

The energy in this region is therefore

$$U_2 = \frac{1}{2} L_2 i^2 = \frac{1}{2}\mu_0 \int H^2\, dv = \frac{\mu_0 i^2 R}{8}$$

and

$$L_2 = \frac{\mu_0 R}{4}$$

Since L_1 was found to be $\mu_0 R\left(\log_e \frac{8R}{r} - 2\right)$, the total self-inductance which is the sum of these two is

$$L = \mu_0 R\left(\log_e \frac{8R}{r} - \frac{7}{4}\right)$$

The self-inductance of a circular coil of n turns with a radius of curvature

large in comparison with its cross-sectional radius is to a good approximation n^2L.

Problems

1. A hydrogen ion enters a uniform magnetic field at right angles to the lines of induction and is deflected in a circular path. What must be the value of B in order that the ion should transverse a complete circle in 10^{-8} sec.?

2. Calculate the natural period of rotation of an electron in the earth's field (taken as 1 gauss). What is the radius of curvature of the path of a 100-volt electron moving normally to this field?

3. A magnetron consists of a filament of 0.2 mm. radius surrounded by a cylindrical plate of 3 cm. radius. It is observed that when the tube is placed with its axis parallel to the lines of induction in a uniform field, the electron current from the filament to the plate drops to zero for potentials less than 8.85 volts. Show that $B = 6.66$ gauss.

4. Under the influence of ultraviolet light electrons are emitted normally with negligible velocities from one plate of a parallel-plate condenser of separation d which is situated in a magnetic field with B parallel to the plates. Show that if V, the potential difference between the plates, is less than $\frac{1}{2}\frac{e}{m}d^2B^2$, no electron current will reach the positive plate.

5. Show that the drift velocity of an ion with a charge-mass ratio e/m in a region of magnetic induction $\mathbf{B}$ and gravitational acceleration $\mathbf{g}$ is given by $m(\mathbf{g} \times \mathbf{B})/eB^2$.

6. Electrons that have fallen through a potential difference V form a uniform cylindrical beam of radius r constituting a current i in a field-free region. Show that the total electrostatic and electromagnetic radial accelerations of a peripheral electron are given by

$$a = \frac{i}{2\pi\kappa_0}\left(\frac{e}{2mV}\right)^{\frac{1}{2}}\frac{1}{r}\left[1 - \left(\frac{v}{c}\right)^2\right]$$

where c is written for $(\mu_0\kappa_0)^{-\frac{1}{2}}$ and v is the electron velocity. Show that the ratio $(v/c)^2$ is less than 0.01 for values of V less than 2,500 volts.

7. Assuming that the trajectories of all ions entering the mass spectrograph of Fig. 9.7 are normal to the plane of the photographic plate, show that the smallest percentage difference in mass that can be detected is the ratio of the width of the entrance slit to the radius of curvature. Making the opposite assumption, that this ratio is negligible, show that the smallest percentage difference in mass that can be resolved is equal to δ^2, where δ is the angle between the extreme and central entering trajectories.

8. A beam of electrons after falling through a potential of 100 volts spreads out in a cone of small apex angle. What current must be sent through a solenoid of 10 turns per centimeter coaxial with the cone in order to refocus the electrons in a distance of 10 cm.?

9. Calculate and plot the magnetic induction, for both internal and external points, as a function of the distance from the center of a long cylindrical wire 2 mm. in diameter carrying a current of 10 amp.

10. Calculate the force of repulsion per meter between two long parallel wires 30 cm. apart carrying currents of 50 amp. in opposite directions.

11. Show that the field inside a toroid of n turns carrying a current i is the same as that which would be produced by a current ni flowing along the axis of symmetry of the toroid.

12. Show that the magnetic induction in webers per square meter at the center of a square circuit of length l on a side carrying a current i is $\frac{2\sqrt{2}\mu_0}{\pi}\frac{i}{l}$, where i is in amperes and l is in meters.

13. A circuit is in the form of a regular polygon of n sides inscribed in a circle of radius a. If it is carrying a current i, show that the magnetic induction at the center is given by $\frac{\mu_0 ni}{2\pi a} \tan \frac{\pi}{n}$. Show that this expression approaches the induction at the center of a circle as n is indefinitely increased.

14. A cylindrical cavity is drilled in a long solid cylindrical conductor. The axis of the cavity is parallel to that of the conductor but is displaced from it a distance $\mathbf{a}$. If $\mathbf{i}'$ is the current density flowing axially through the conductor, show that the magnetic induction in the cavity is uniform and equal to $\mu_0(\mathbf{i}' \times \mathbf{a})/2$.

15. Show that the axial magnetic field due to a circular current can be written $\frac{1}{2}\frac{i}{b}\sin^3 \theta$, where b is the radius of the circuit and θ is the angle subtended by a radius at the axial point.

16. A sphere of radius a is charged to a uniform surface density σ and rotated about an axis through its center with an angular velocity ω. Show that the induction at the center is given by $\frac{2}{3}\mu_0 \sigma a \omega$ along the axis of rotation. Show that the induction has the same value at any point on the axis. Show by means of the equation curl $\mathbf{H} = 0$ inside the sphere that this is the value of the induction throughout the spherical volume.

17. A fine insulated wire is wound in a close spiral forming a circular disk, the ends being at the center and the circumference. Show that the induction at a point on the axis of this disk is given by

$$\tfrac{1}{2}\mu_0 ni[\cosh^{-1} (\sec \theta) - \sin \theta]$$

where i is the current flowing in the wire, n is the number of turns per unit radius, and θ is the angle subtended by a radius of the disk at the point on the axis.

18. A wire is wound in a helix of angle α on the surface of an insulating cylinder of radius a so that it makes n complete turns on the cylinder. If a current i flows in the wire, show that the resultant magnetic induction at the center of the cylinder is

$$\frac{1}{2}\mu_0\frac{in}{a}(1 + \pi^2 n^2 \tan^2 \alpha)^{-\frac{1}{2}}$$

19. Show that the lines of magnetic induction produced by equal and opposite currents flowing in long parallel filaments a distance d apart are the same as the equipotentials in the case of long parallel cylindrical conductors oppositely charged, *i.e.*, circles of radius $nd/(n^2 - 1)$ with their centers displaced a distance $d/(n^2 - 1)$ on the far side of each filament from the axis of symmetry.

20. The order of magnitude of the radius of an electron can be obtained by considering that its mass is entirely electromagnetic in origin. This means that its kinetic energy ($\frac{1}{2}mu^2$) is equal to that calculated from Eq. (9.27) with a for the lower radial limit of integration. Show that the electron radius calculated on this basis is $\mu_0 e^2/6\pi m$ or approximately 2×10^{-15} m.

21. Show that the torque on an elementary vortex of current i having an area a (moment $\mathbf{m} = \mu_0 i\mathbf{a}$) in a magnetic field $\mathbf{H}$ is given by $\mathbf{m} \times \mathbf{H}$. Show that the force

on such an elementary circuit can be written (**m** · **grad**) **H** by which is meant the vector having an x component

$$m_x\frac{\partial H_x}{\partial x} + m_y\frac{\partial H_x}{\partial y} + m_z\frac{\partial H_x}{\partial z}$$

and analogous expressions for the y and z components.

22. A long flexible copper helix of radius b and negligible weight has n' turns per meter. It is suspended from its upper end and to its lower end is attached a mass m resting on a metal table. Assuming that the solenoid contracts uniformly, show that the current that must be sent through it in order to lift the mass off the table is given by $\frac{1}{n'b}\sqrt{\frac{2mg}{\pi\mu_0}}$.

23. A string galvanometer consists of a thin wire held under a tension T by means of a helical spring and set vertically in a horizontal magnetic field. A current i flows in the wire and the horizontal displacement d of the center of the wire is measured with a microscope. Show that the wire assumes the form of a parabola and that d is given by $il^2B/8T$, where B is the induction and l is the distance between the points of support of the wire.

24. Show that if a complex alternating-current wave is applied to a dynamometer, the displacement of the pointer is a measure of $\frac{1}{2}\sum_{j=1}^{j=n} a_j^2$, where a_j is the amplitude of the jth harmonic.

25. Show that a circular loop of wire carrying a current i will expand radially if T is less than $\frac{i^2}{4\pi}\frac{\partial L}{\partial R}$, where T is the breaking tension of the wire and L is the self-inductance of the loop.

26. Two coaxial coils of radius a and separated by a distance a (Helmholtz coils) produce an approximately uniform magnetic field in the region near the axis and halfway between them. Show that the axial field in this central region is given by $\left(\frac{4}{5}\right)^{3/2}\frac{ni}{a}$ to terms of the order $(x/a)^2$, where x is the distance along the axis from the point halfway between the coils, and n and i are the number of turns and the current in each coil, respectively.

27. A solenoid 50 cm. long and 2 cm. in diameter contains 1,000 turns. It is over-wound near its center with a second coil of 100 turns. What is the mutual inductance between the coils?

28. Show that the current distribution due to the spinning sphere of Prob. 16 is equivalent to a system of coaxial circular circuits coinciding with the spherical shell and carrying a current $\sigma\omega a$ per unit length along the axis and hence the sphere has a magnetic moment equal to $\mu_0\omega a^2 e/3$, where e is the total charge.

29. Assuming a homogeneous isotropic medium of conductivity σ, show by means of Eq. (9.18) that for steady currents **H** satisfies Laplace's equation.

30. In a mass spectrometer, of the type for example shown in Fig. 9.7, the beam of ions spreads apart owing to the repulsion of the charges of like sign composing the beam. Assuming that Fig. 9.7 represents a two-dimensional apparatus in which the beam is a ribbon extending indefinitely normal to the page and that the radius of curvature ρ is very large compared with the width of the beam, show that the focal line of the beam will be spread to a width w by space charge after traversing the angle π

where

$$w = \frac{1}{2}\frac{i\pi^2}{\kappa_0 \rho B^3}\left(\frac{m}{q}\right)^2$$

Here i is the beam current per unit length normal to the page.

31. A wedge-shaped region of uniform magnetic field is bounded by two planes that meet at the angle φ in a line L parallel to the field. A beam of ions of small conical divergence leaves point P_1, and the axis of the beam enters the region of magnetic field normal to one of the wedge surfaces at a distance ρ from L where ρ is the radius of curvature of the ions in the magnetic field. Show that the beam on leaving the other wedge surface will be brought to an astigmatic focus on a line through the point P_2 where P_2 lies in the plane containing P_1 and L. Show also that the following lens-type formula relates the distances $d_1 = P_1L$ and $d_2 = LP_2$

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f}$$

where $f = \rho/2 \sin(\varphi/2) \cos(\varphi_1 - \varphi/_2)$ and φ_1 is the angle between the axis of the initial beam and the line P_1P_2.

32. A segment of the electric field between two coaxial cylindrical condenser plates is bounded by two planes meeting at the axis with included angle φ. A divergent beam of ions of small conical angle α leaves normally from a point on one of these planes. The velocity and sign of charge are such that the ions on the axis of the cone describe a circle of constant radius r_0 about the cylindrical axis of the field. Determine the necessary condition between the energy of the ions and the electric field at r_0 for this to occur, and show that the beam is brought to an astigmatic focus after traversing an angle $\pi/\sqrt{2}$ about the axis of the cylinder.

33. Show that Eq. (9.16) for the field due to a dipole of moment $\mathbf{m}$ at some distance r can be written as

$$\mathbf{H} = \frac{1}{4\pi\mu_0}\,\mathbf{grad}\left(\mathbf{m}\cdot\mathbf{grad}\left(\frac{1}{r}\right)\right)$$

Show that the magnetic vector potential $\mathbf{A}$ due to the dipole can by the aid of the vector identities (Appendix D) be written

$$\mathbf{A} = \frac{1}{4\pi}\mathbf{m} \times \mathbf{grad}\left(\frac{1}{r}\right) = \frac{1}{4\pi}\,\mathbf{curl}\left(\frac{\mathbf{m}}{r}\right)$$

CHAPTER X

CHANGING ELECTRIC CURRENTS AND ELECTROMAGNETIC REACTIONS

10.1. Faraday's Law of Induction.—The previous chapter was principally concerned with the mechanical forces between circuits carrying steady currents or charges moving with uniform velocities. Ampère's law of force was used to construct an energy function which was written in the various forms of Eq. (9.24) to (9.27). It is a function of the currents and the variables necessary to specify the positions of the circuits. As the currents were considered constant, the forces or torques were derived by taking the positive derivatives of the energy function with respect to the appropriate variable. It was found both by Faraday and Henry about 1831 that electrical as well as mechanical forces are developed when the currents or the relative positions of the circuits are altered. The electrical forces are manifested by changes that are induced in the circulating currents. Thus, if there are two circuits in close proximity one of which contains a galvanometer and the other a battery, when the battery circuit is made and then broken, the galvanometer is observed to deflect first in one direction and then in the other. While the current is being established in the battery circuit, an electromotive force is induced in the galvanometer circuit, which causes a current to flow, and when the battery circuit is opened so that the current ceases to flow, an oppositely directed electromotive force is induced in the galvanometer circuit producing a flow of current in the reverse direction. There is no electromotive force in the galvanometer circuit except when the current in the battery circuit is changing. If the relative sense of the current in the galvanometer circuit is examined, it is found that this current flows in such a direction as to counteract the change in magnetic induction caused by the change in the current flowing in the battery circuit. Likewise it is found that if the key in the battery circuit is permanently depressed and one of the circuits is moved relative to the other, a deflection of the galvanometer is observed.

Fig. 10.1.—The induction of currents.

That is, a change in the relative position of the circuits when one of them carries a current induces an electromotive force just as a change in the current does. The electromotive force induced by a change in the relative position of the circuits is sometimes known as a *motional electromotive force.* Its sense can be described in the same way as that due to a changing current. It may be stated in general that any change in the flux of magnetic induction through a circuit induces an electromotive force around it in such a sense as to give rise to a current opposing the change. This is known as *Lenz's law.* It may be considered as the electrical analogue of Newton's law of action and reaction or as an electrical application of the general principle of Le Chatellier, which states that if any state of equilibrium is disturbed, a reaction is induced in such a sense as to oppose the change.

Faraday's law of induction may be considered to be an experimental law as basic for electromagnetism as that of Ampère. Alternatively it may be shown to be a consequence of the principle of the conservation of energy on the assumption that the expressions for the energy of the previous chapter, such as Eq. (9.24) which was derived for the case of constant currents only, are equally valid as expressions for the magnetic energy when the currents are permitted to vary. In the following discussion it is assumed that the current-carrying wires are fixed in shape and position so that no mechanical work is performed. If the total energy U is to remain constant for variations in the currents, any change in U implied by the current changes must be counterbalanced by work done by electrical forces operating in the circuits. Let δi_j be the change of the current in the j'th circuit and δU the change in magnetic energy. On applying Taylor's theorem to Eq. (9.24), the change in energy can be written

$$\delta U = \sum_k i_k \left\{ \sum_j L_{jk}\, \delta i_j \right\}$$

Since the flux through circuit k is by definition $\sum_j L_{jk} i_j$, $\sum_j L_{jk} \delta i_j$ is the change in flux through this circuit or $\delta\phi_k$. Thus the change in energy can be written

$$\delta U = \sum_k i_k\, \delta\phi_k$$

Passing to infinitesimals and assuming that the current changes take place in an infinitesimal time dt, the rate of change of magnetic energy can be written

$$\frac{\partial U}{\partial t} = \sum_k i_k \frac{\partial \phi_k}{\partial t}$$

As the circuits are at rest, no mechanical work is done, and hence if energy is to be conserved, this must be equal to minus the rate at which the electrical forces are doing work. If $\mathcal{E}_1$ is the induced electromotive force in the kth circuit, the rate of working of the electrical forces in that circuit is $\mathcal{E}_k i_k$ and hence

$$\sum_k i_k \frac{\partial \phi_k}{\partial t} = -\sum_k \mathcal{E}_k i_k$$

Since the currents in the circuits are all independent, corresponding terms in the summations must be separately equal to one another or

$$\mathcal{E}_k = -\frac{\partial \phi_k}{\partial t} \tag{10.1}$$

Of course, there will in general be batteries or generators producing emfs. $\mathcal{E}'_k$ which maintain the currents i_k in the various circuits and which dissipate electrical energy at the rate $\mathcal{E}'_k i_k$, but $\mathcal{E}_k$ is an additional emf. which is produced by the interchange of energy from the magnetic to the electric form.

Equation (10.1) for the induced electromotive force is known as *Faraday's law of induction.* As in the cases of the magnetic field and the magnetic energy, this equation also can be written in various forms that are useful in different circumstances. In dealing with stationary wire circuits it is more convenient to express the flux in terms of the coefficients of induction. Since $\phi_k = \sum_j L_{jk} i_j$

$$\mathcal{E}_k = -\sum_j L_{jk} \frac{di_j}{dt}$$

The change in the current is written as a total derivative since the use of coefficients of induction assumes that the currents are uniform throughout the circuits and the only variable appearing is the time. In the case of a single circuit of self-inductance L, a change in the current gives rise to an induced or inertial electromotive force $L\frac{di}{dt}$ which opposes the growth or decay of a current. If V is the potential difference applied to a linear circuit, $V - L\frac{di}{dt} = Ri$, or

$$L\frac{di}{dt} + Ri = V \tag{10.2}$$

This is the general linear differential equation for a circuit containing inductance and resistance. It is of the form of Eq. ($C.5$) (Appendix C)

and the general solution is given by Eq. (C.6). However, the cases that are most frequently encountered are those in which V is either a constant or a simple periodic function. In the former case the variable can be changed to $i' = i - \frac{V}{R}$ and the equation written

$$L\frac{di'}{dt} + Ri' = 0$$

The variables are separable and this equation can be integrated immediately to yield

$$i' = i'_0 e^{-\frac{Rt}{L}}$$

where i'_0 is the value of i at $t = 0$. Thus, if $i = 0$ at $t = 0$

$$i = \frac{V}{R}(1 - e^{-\frac{Rt}{L}})$$

This represents the growth of a current as shown in Fig. 10.2. If $i = V/R$ at $t = 0$ and subsequently $V = 0$, then

$$i = \frac{V}{R}e^{-\frac{Rt}{L}}$$

This represents the decay of the current when the battery of potential V is shorted. L/R, which is the time for the current to fall to $1/e$ of its

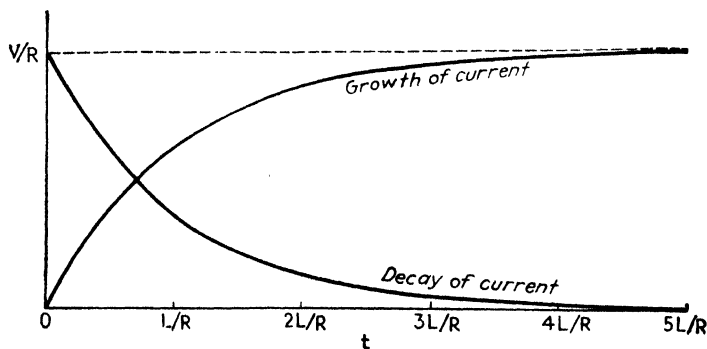



FIG. 10.2.—Growth and decay of a current in an inductive circuit.

initial value, is called the *time constant* of the circuit. The case in which V is a periodic function of the time can be handled exactly as the resistance-capacity circuit of Sec. 7.6. Writing $\mathbf{V}$ for the complex potential $V_0 e^{j\omega t}$ and assuming $\mathbf{i} = i_0 e^{j\omega t}$, Eq. (10.2) becomes

$$i_0(R + j\omega L)e^{j\omega t} = V_0 e^{j\omega t}$$

or

$$\mathbf{i} = \frac{\mathbf{V}}{\mathbf{z}}$$

where the complex impedance $\mathbf{z}$ is $R + j\omega L$. The term ωL is known as the *inductive reactance*. The absolute magnitude of the impedance is $(R^2 + \omega^2 L^2)^{1/2}$ and this is the ratio of the amplitude of the potential wave to the amplitude of the current wave. Because of the linear relation between **i** and **V** the current and potential waves are of the same form, but the former lags behind the latter by the phase angle φ which is given by $\tan^{-1}(\omega L/R)$, or by a time φ/ω. The current and potential waves in a resistance-inductance circuit are shown in Fig. 10.3, which may be compared with the resistance-capacity circuit of Fig. 7.21. The general alternating-current theory of circuits containing resistance, capacity, and inductance will be postponed till Chap. XIII.

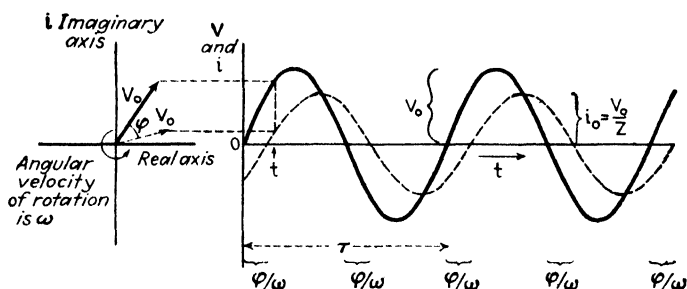



FIG. 10.3.—Potential and current waves in a resistance-inductance circuit.

10.2. Induction of Currents in Continuous Media.—For a discussion of the electric and magnetic vectors in a continuous medium or in free space it is more convenient to write Eq. (10.1) in a different form. The electromotive force around a closed circuit is by definition the integral of $\mathbf{E} \cdot \mathbf{dl}$ around the circuit, or dropping the subscripts Eq. (10.1) can be written

$$\oint \mathbf{E} \cdot \mathbf{dl} = -\frac{\partial \phi}{\partial t}$$

where ϕ is the total flux enclosed by the path of integration. It is unnecessary that the path of integration should coincide with a conducting wire. This is evident from the fact that the tangential component of **E** is continuous across the boundary between a wire and the space outside it. Thus, if the equation is true for a path just inside the wire, the integral $\oint \mathbf{E} \cdot \mathbf{dl}$ has the same value for a neighboring path just outside. Hence Eq. (10.1) can be taken as true for any path whatever. Now, by definition

$$\phi = \int \mathbf{B} \cdot \mathbf{ds}$$

where the surface over which **B** is summed is any surface bounded by the path around which the integral $\oint \mathbf{E} \cdot \mathbf{dl}$ is taken. And by Stokes's

theorem [Eq. (D.15) Appendix D]

$$\oint \mathbf{E} \cdot \mathbf{dl} = \int \mathbf{curl\ E} \cdot \mathbf{ds}$$

Hence

$$\int \mathbf{curl\ E} \cdot \mathbf{ds} = -\frac{\partial}{\partial t}\int \mathbf{B} \cdot \mathbf{ds}$$

or constricting the surface of integration to an infinitesimal one

$$\mathbf{curl\ E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{10.3}$$

This is the differential equation relating the electric field and magnetic induction at any point which forms the basis for the discussion of the induction of currents in continuous media.

The differential equations for the electric and magnetic vectors can be obtained separately by means of Eq. (9.18). Multiplying by the conductivity σ, which is assumed to be constant, and taking the **curl** of Eq. (10.3)

$$\mathbf{curl\ curl}\ \sigma\mathbf{E} = -\sigma\frac{\partial}{\partial t}\ \mathbf{curl\ B}$$

We have also the general vector equation

$$\mathbf{curl\ curl\ i}_v = \mathbf{grad}\ \mathrm{div}\ \mathbf{i}_v - \nabla^2\mathbf{i}_v$$

And it will be recalled that $\mathbf{i}_v = \sigma\mathbf{E}$. Assuming for the moment that there is no accumulation of charge so that div $\mathbf{i}_v = 0$ and writing $\mu_0\mathbf{H}$ for $\mathbf{B}$

$$-\nabla^2\mathbf{i}_v = -\sigma\mu_0\frac{\partial}{\partial t}\ \mathbf{curl\ H}$$

On using Eq. (9.18) to eliminate the magnetic vector, the equation for the current in the conducting medium can be written

$$\nabla^2\mathbf{i}_v = \sigma\mu_0\frac{\partial \mathbf{i}_v}{\partial t} \tag{10.4}$$

Induced currents of this type are known as *Eddy currents*. The current can be eliminated instead of the magnetic vector and it is then found that the induction **B** satisfies the same differential equation as the current. This will be recognized as being the same differential equation as that describing the conduction of heat. The quantity $1/\sigma\mu_0$ is a measure of the rate at which the currents diffuse into the conductor. In order to obtain a description of the induced currents in any particular case, a solution of the equation must be obtained which satisfies the imposed boundary conditions.

As an example of the use of this equation consider that the inducing flux varies in a simple periodic manner. Since the equation is linear, the current will also vary in the same way with the time and $\mathbf{i}_v$ can be written $\mathbf{i}'_v e^{j\omega t}$. On this assumption Eq. (10.4) becomes

$$\nabla^2 \mathbf{i}'_v = j\omega\mu_0\sigma\mathbf{i}'_v$$

Let us take the special case of a semiinfinite conducting medium bounded by the plane $z = 0$ and assume that the currents circulate parallel to this plane, $\mathbf{i}'_v$ being a function only of z, the distance in from the boundary. Then writing $\mathbf{i}$ for $\mathbf{i}'_v$

$$\frac{d^2\mathbf{i}}{dz} = j\omega\mu_0\sigma\mathbf{i}$$

A particular solution of this equation is $i = i_0 e^{\alpha z}$. The constant α is determined by substituting this value of $\mathbf{i}$ in the equation, yielding

$$\alpha^2 = j\omega\mu_0\sigma$$

or since the square root of j is $(1 + j)/\sqrt{2}$

$$\alpha = \pm\left(\frac{\omega\mu_0\sigma}{2}\right)^{1/2}(1 + j)$$

The negative sign must be chosen in order that the current shall not become infinite with increasing z. The final equation for $\mathbf{i}_v$ is then

$$\mathbf{i}_v = \mathbf{i}_0 e^{-\sqrt{\frac{\omega\mu_0\sigma}{2}}z} e^{j\left(\omega t - \sqrt{\frac{\omega\mu_0\sigma}{2}}z\right)} \tag{10.5}$$

where $\mathbf{i}_0$ is the current just inside the boundary. The second exponential term represents the periodic variation of the current, the second term in the bracket being the phase angle of the current at a depth z with respect to the surface current $\mathbf{i}_0$.

The first exponential term indicates the damping or decrease in amplitude with increasing depth z. The damping is seen to increase with the angular frequency; this is analogous in thermal conduction to the fact that diurnal fluctuation of temperature is not appreciable as far below the surface of the earth as the seasonal variation. To obtain the order of magnitude of the damping factor for currents in a good conductor consider copper, for which the value of σ is 5.8×10^7 mho per meter. In this case the exponential radical is approximately $6\sqrt{\omega}$ per meter. For a 60-cycle fluctuation the coefficient of z is approximately 120 so that, for instance, at a depth of 5 cm. the currents are less than 1 per cent of their value at the surface. This tendency for induced currents to flow near the surface becomes very important at radio frequencies. At 60 megacycles, for instance, the currents would be less than 1 per cent of their surface value at a depth of 0.05 mm. Thus the currents are largely confined to a thin layer of the metal immediately

below the surface, and the phenomenon is known as the *skin effect*. Since the cross section through which the current flows is much less than the total area of the conductor the effective resistance increases with the frequency. Thin-walled tubes can be used just as efficiently as solid conductors at very high frequencies. Since the magnetic flux obeys the same equation as the current, the induction inside the conductor decreases as the frequency increases. This means that the magnetic energy in the conductor is less than for steady currents and, as can be seen from the methods of calculating the coefficients of inductance in Sec. 9.6, this implies that the self-inductance of a circuit can be expected to decrease slightly as the frequency increases. These results will be considered further in connection with other high-frequency phenomena.

For certain magnetic materials such as iron, which will be considered in the following chapter, the effective value of μ_0 may be several hundred times the value for free space. Assuming a value of about $600\mu_0$ and taking σ to be about $\frac{1}{6}$ the value for copper, it is seen from Eq. (10.5) that the circulating currents at 60 cycles are less than 1 per cent of their surface value at a depth of about 0.5 cm. Thus for iron the currents and induction are still more closely confined to the surface of the conductor. An extreme instance of the suppression of volume currents is found in the case of metals in the superconducting state.[1]

Though our present understanding of superconductivity in terms of the electron and band structure of crystals is far from complete there appears to be a very intimate connection between it and diamagnetic permeability (Sec. 11.2). When, for instance, lead is cooled below its critical temperature of about 4°A., its conductivity is found to be about 10^{12} times its conductivity at room temperature, *i.e.*, of the order of 5×10^{18}. If this cooling takes place in a moderatė magnetic field it is found that the lines of induction become distorted with the onset of superconduction in such a way that they no longer enter the block of lead. Those previously traversing it are forced outside and the magnetic induction within the block is effectively zero. If the lead block contains a cavity the induction in the cavity may be altered somewhat but remains comparatively unaffected below the superconduction temperature. If a moderate magnetic field is applied to the lead after it has been cooled below the critical temperature for superconduction, again no lines of induction enter the lead block but they are all deflected around the outer boundary. The existence of a magnetic field lowers the critical temperature at which a material becomes superconducting. If the field exceeds, say, 10^{-2} or 10^{-1} webers/meter^2 the phenomena of superconduc-

[1] For summaries of the phenomena exhibited by superconductors in magnetic fields see Smith and Wilhelm, *Rev. Mod. Phys.* **7**, **237** (1935); Collins, *Science*, **107**, **327** (1948).

tion may never appear. Since a magnetic field accompanies an electric current a sample which is superconducting for low currents will no longer be superconducting if the current exceeds a value for which the associated magnetic field would inhibit the occurrence of superconductivity.

As the magnetic induction is zero within a superconductor the currents are limited to the surface of a sample. Experiments indicate that the surface layer in which currents flow has a thickness of the order of 10^{-7} meters. It is unlikely that Eq. (10.5) applies directly to a superconductor, but taking σ to be 5×10^{18} the coefficient of z in the first exponent becomes $1.8 \times 10^6\sqrt{\omega}$ indicating a penetration depth of the order of that found experimentally. There is some variation of the thickness of the superconducting layer with frequency but the precise nature of it has yet to be experimentally established. As the superconduction currents induced by a magnetic field are such as to effectively preclude the existence of any induction within the material, the region within a superconductor or any volume completely surrounded by superconducting walls is completely shielded from any external changes in the magnetic induction. Ordinary conductors exert a partial shielding effect of the same nature, but since the currents induced in them decrease with time owing to their finite resistivity, the establishment of a magnetic field in a region surrounded by such a conductor and also in neighboring external regions merely lags in time behind the imposed field. However, for sufficiently rapid fluctuations the penetration of the field in ordinary conductors becomes very small and the internal region is almost completely shielded from external magnetic effects.

FIG. 10.4.—Displacement current through a condenser.

Before leaving this discussion it is necessary to clear up a point that arose earlier in this section in connection with the divergence of the current density. It was there assumed that div $\mathbf{i}_v = 0$, which is in agreement with Eq. (9.18), for on taking the divergence of both sides of this equation, since div **curl** $\mathbf{H} = 0$, identically, div $\mathbf{i}_v$ must also be identically zero. However, it is possible for charges to accumulate in a region as, for instance, on the plates of the condenser illustrated in Fig. 10.4. When the key is closed in the battery circuit, an electromotive force is induced in the circuit, containing the condenser, and electrons are circulated from one plate of the condenser to the other, producing a positive charge on one and a negative charge on the opposite one. Considering the volume bounded by the dashed surface enclosing one plate, it is seen that charges accumulate in this region at a rate i or dq/dt. Proceeding to the general case,[1] the conservation of charge implies that

[1] See Sec. 3.4.

the total outward normal flow of charge over any closed surface must equal the rate at which the total enclosed charge is decreasing. Or since the surface integral of the normal component of the current density is equal to the volume integral of its divergence

$$\int \operatorname{div} \mathbf{i}_v \, dv = -\int \frac{\partial q_v}{\partial t} dv$$

On constricting the integration to an infinitesimal volume

$$\operatorname{div} \mathbf{i}_v = -\frac{\partial q_v}{\partial t} \tag{10.6}$$

Since the right-hand side of this equation is not in general zero, it is in contradiction with Eq. (9.18).

This difficulty was resolved by Maxwell who made the assumption that the conduction current in Eq. (9.18) must be supplemented by another term to obtain a total effective current the divergence of which will vanish. On writing this additional contribution to the current as $\mathbf{i}_d$, Eq. (9.18) becomes

$$\mathbf{curl\ H} = \mathbf{i}_v + \mathbf{i}_d$$

and it is necessary that $\operatorname{div}(\mathbf{i}_v + \mathbf{i}_d) = 0$. Now, on comparing the general equation of electrostatics $\operatorname{div} \mathbf{D} = q_v$ [Eq. (2.21)] with Eq. (10.6) it is evident that

$$\frac{\partial}{\partial t} \operatorname{div} \mathbf{D} = -\operatorname{div} \mathbf{i}_v$$

or neglecting unimportant uniform fields or currents the conduction current is given by

$$\mathbf{i}_v = -\frac{\partial \mathbf{D}}{\partial t}$$

In order that $\operatorname{div}(\mathbf{i}_v + \mathbf{i}_d)$ shall be identically zero, it is necessary that $\mathbf{i}_v = -\mathbf{i}_d$ or $\mathbf{i}_d = \partial \mathbf{D}/\partial t$. Thus, when variations with the time are considered, Eq. (9.18) must be written

$$\mathbf{curl\ H} = \mathbf{i}_v + \frac{\partial \mathbf{D}}{\partial t} \tag{10.7}$$

The introduction of $\mathbf{i}_d$, which is known as the *displacement current*, has rendered the equations consistent. It was originally interpreted by Maxwell as representing a change in the state of strain of the ether which was responsible for the transmission of electric and magnetic forces through space. It is very improbable that such a medium exists and the physical interpretation of the displacement current is not at all obvious. However, it may be looked upon formally as a property of

space itself. As will be seen in Chap. XVI, this concept has very important consequences for insulators and for free space where the conduction current vanishes. However, for conductors the difference between Eqs. (9.18) and (10.7) is quite unimportant and the previous results are essentially valid. This may be seen by considering the ratio $\mathbf{i}_d/\mathbf{i}_v$. Since $\mathbf{D} = \kappa\kappa_0\mathbf{E}$ and $\mathbf{i}_v = \sigma\mathbf{E}$

$$\frac{\mathbf{i}_d}{\mathbf{i}_v} = \frac{\kappa\kappa_0\, \partial\mathbf{E}/\partial t}{\sigma\mathbf{E}}$$

Considering for instance a periodic field $\mathbf{E} = \mathbf{E}_0 e^{j\omega t}$, the magnitude of this ratio becomes $\kappa\kappa_0\omega/\sigma$. Taking the case of copper for which $\sigma = 5.8 \times 10^7$ and inserting the value 8.85×10^{-12} for κ_0, the ratio $\kappa\kappa_0/\sigma$ is of the order of 10^{-19}. Hence $\mathbf{i}_d$ is less than 1 per cent of $\mathbf{i}_v$ for frequencies less than 10^{16} per second. This frequency is in the range of ultraviolet light and larger by a factor of 10^6 than any that are encountered in electrical circuit work; hence the neglect of the displacement current in any conducting medium is entirely justified.

10.3. Motional Electromotive Force.—It was mentioned earlier in this chapter that an electromotive force is induced in a circuit which is moving relative to a magnetic field. A complete description of the forces induced by the relative motion of charges requires the use of the theory of relativity. An account of the relativity theory is beyond the scope of this treatment, and for a discussion of the interaction of moving charges from this point of view reference should be made to more mathematical treatises.[1] However, correct results for velocities which are small in comparison with that of light can be obtained on the basis of the simple electron theory that has been used in the previous discussions. Equation (9.6) implies that when a charge moves with a velocity $\mathbf{u}$ in a region of magnetic induction $\mathbf{B}$, a supplementary electric field equal to $\mathbf{u} \times \mathbf{B}$ is effectively brought into existence. The physical existence of this induced electric field has been shown by an experiment of Wien,[2] in which he observed that the radiation from an atom moving relative to a magnetic field is modified in exactly the same way as it is found to be by the presence of an electric field of magnitude $\mathbf{u} \times \mathbf{B}$. Thus, when a conductor such as a copper wire containing free conduction electrons is moved in a magnetic field, the electrons circulate under the influence of this field and in the case of a wire they are constrained by the boundary of the wire to move principally in the direction of its length. If an

[1] Electrodynamics of Moving Media, *Nat. Res. Council Bul.* (1922); FRENKEL, "Lehrbuch der Elektrodynamik," Berlin, 1926; SMYTHE, "Static and Dynamic Electricity," McGraw-Hill Book Company, Inc., New York, 1939; STRATTON, "Electromagnetic Theory," McGraw-Hill Book Company, Inc., New York, 1941.

[2] WIEN, *Ann. Physik*, **81**, 994 (1926).

extended conductor moves in a field, the electrons circulate throughout the body of the conductor. By Lenz's law the currents to which they give rise are in such a sense as to oppose the changing flux and hence to oppose the motion of the conductor. This is also obvious from the conservation of energy since if these induced currents aided the motion, a small displacement would give rise to an acceleration in the same direction and the momentum would continue to increase spontaneously without the expenditure of any external work. This retarding force that is induced by the motion of a conductor in a magnetic field is known as an *electromagnetic reaction.*

Though this electromotive force is presumably brought into existence only if there are charges moving in a region of magnetic induction, it may be considered to be essentially the same as that produced by a changing flux. By Stokes's theorem

$$\mathcal{E}' = \oint \mathbf{E}' \cdot d\mathbf{l} = \oint (\mathbf{u} \times \mathbf{B}) \cdot d\mathbf{l} = \int [\mathbf{curl}\ (\mathbf{u} \times \mathbf{B})] \cdot d\mathbf{s}$$

where the surface of integration is bounded by the closed path of integration, $\oint$, lying completely in a conductor. The total emf. produced both by a changing flux and by the motion of a conductor can be written with the aid of the last equality and Eq. (10.3) as

$$\mathbf{curl\ E} = -\left[\frac{\partial \mathbf{B}}{\partial t} - \mathbf{curl}\ (\mathbf{u} \times \mathbf{B})\right]$$

The right-hand side of the equation can be considered as the negative total derivative of **B** with respect to t for a rigid moving conductor which is the instance most frequently encountered. For on expanding **curl** (**u** × **B**) with the understanding that **u** is not a function of the spacial variables and recalling that div **B** = 0, it is seen to reduce to the vector − (**u** · **grad**)**B**.* The x component of this vector is

$$-\frac{\partial B_x}{\partial x}u_x - \frac{\partial B_x}{\partial y}u_y - \frac{\partial B_x}{\partial z}u_z$$

and analogous expressions for the other components. If **B** is a function of t and the spacial variables

$$\frac{dB_x}{dt} = \frac{\partial B_x}{\partial t} + \frac{\partial B_x}{\partial x}\frac{dx}{dt} + \frac{\partial B_x}{\partial y}\frac{dy}{dt} + \frac{\partial B_x}{\partial z}\frac{dz}{dt}$$

Thus the right-hand side of the equation for **curl E** is seen to be the same as the total derivative of **B** with respect to t, and Eq. (10.3) can be generalized to include the motion of rigid conductors as

$$\mathbf{curl\ E} = -\frac{d\mathbf{B}}{dt} \tag{10.8}$$

* Appendix *D*.

or by Eq. (9.12)

$$\mathbf{E} = -\frac{d\mathbf{A}}{dt} \tag{10.8'}$$

Hence the apparent electric field that is detected by the force exerted on a charge is the total negative time derivative of the vector potential. Also by integrating Eq. (10.8) around a closed circuit it is seen that the emf. is equal to the total rate of decrease of flux through the circuit. Faraday's law of induction and Ampère's law of force are the two principles underlying the operation of all moving electromagnetic machinery, and various special instances will be encountered in the subsequent discussions.

The general methods of dynamics may be used to solve problems involving the motion of charges on the basis of forces arising from the laws of Coulomb, Ampère, and Faraday. Allowing for the possibility of forces of a nonelectrical nature that can be described in terms of the gradient of a potential function P, the vector equation for the motion of a mass m having a charge q is

$$\mathbf{F} = m\frac{d^2\mathbf{r}}{dt^2} = -\operatorname{\mathbf{grad}} P - q \operatorname{\mathbf{grad}} V - q\frac{\partial \mathbf{A}}{\partial t} + q(\mathbf{u} \times \operatorname{\mathbf{curl}} \mathbf{A})$$

where $\mathbf{u} = d\mathbf{r}/dt$. The Lagrangian form of the mechanical equations of motion is

$$\frac{d}{dt}\left(\frac{\partial \mathsf{L}}{\partial \dot{q}_k}\right) - \frac{\partial \mathsf{L}}{\partial q_k} = 0$$

where the q_k's are the coordinates describing the position of m and the dots represent differentiation with respect to time. If the following expression is chosen for the Lagrangian function L, the equations of motion are consistent with the vector equation above:

$$\mathsf{L} = \tfrac{1}{2}m\dot{r}^2 - P - qV + q\dot{\mathbf{r}} \cdot \mathbf{A}$$

This may also be put in the Hamiltonian form in which the equations of motion are the first-order equations

$$\frac{\partial q_k}{\partial t} = \frac{\partial \mathsf{H}}{\partial p_k} \qquad \text{and} \qquad \frac{\partial p_k}{\partial t} = -\frac{\partial \mathsf{H}}{\partial q_k}$$

where p_k is the momentum conjugate to q_k and is defined as $p_k = \partial\mathsf{L}/\partial\dot{q}_k$. The Hamiltonian function is written explicitly in terms of the q_k's and p_k's and is $\mathsf{H} = \sum_k p_k\dot{q}_k - \mathsf{L}$. From the expression for L, $p_k = m\dot{r}_k + qA_k$, where the subscripts on $\mathbf{r}$ and $\mathbf{A}$ indicate the components of these vectors on the direction of the coordinate q_k. Adopting Cartesian coordinates x_1, x_2, and x_3; $p_{x_i} = m\dot{x}_i + qA_{x_i}$ etc., and

$$\mathsf{H} = \frac{1}{2m}\sum_{i=1}^{i=3}(p_{x_i} - qA_{x_i})^2 + P + qV$$

P, V, and A are all in general functions of the x's.

10.4. Absolute Determination of the Ohm and Measurement of the Constant κ_0.—It was mentioned in connection with the absolute determination of the ampere by means of the current balance that a standard of resistance can be established independently and these two standards used to calibrate a standard cell as a secondary standard of emf. One of the methods of determining the resistance of a chosen length of wire suitably mounted to provide a satisfactory standard is that due to Lorenz. A brass disk is mounted on an axle parallel to a magnetic field and rotated at a uniform angular velocity ω. One electrical contact is made directly to the axle and the other to the periphery by means of a phosphor-bronze spring that presses against the edge of the disk as it rotates. Since the motion of the conductor is in the direction of the angular coordinate θ, which is perpendicular to the induction **B**, a radial emf., perpendicular both to **B** and the velocity, is established between the periphery of the disk and the axle. This emf. is balanced against the iR drop in the standard resistance which is included in the circuit of the coils producing the magnetic field. The arrangement is illustrated schematically in Fig. 10.5. The disk D is rotated by means of the motor M and ν, the number of rotations per second is measured stroboscopically. The two coils A and A' symmetrically placed on either side of the disk carry a current i which produces an axial field through the disk. Since the emf. produced between the axle and periphery of the disk is small, it is balanced against only a fraction of the drop in potential across the resistance R. This is accomplished by means of the resistance network R-R'-R''. If i is the current flowing to the network and a balance is achieved so that no current flows through the galvanometer the potential drop across the resistance R' is $iRR'/(R + R' + R'')$. These resistances are accurately compared with one another and the ratios $R'/R = x'$ and $R''/R = x''$ determined so that this potential drop can be written $iRx'/(1 + x' + x'')$. If L is the mutual inductance between the coils and the disk, the flux through the disk is equal to iL. Since any radius of the disk passes through this flux ν times a second, the emf. induced between the axle and periphery is given by $Li\nu$. If the velocity of rotation of the disk is varied until on tapping the key K no galvanometer deflection is observed, these two emfs. can be equated and the resistance R written explicitly as

FIG. 10.5.—Lorenz method of determining the value of the ohm.

$$R = L\nu\frac{(1 + x' + x'')}{x'}$$

The effects of stray magnetic fields and thermal and contact emfs. can be eliminated by making a second determination of the balance condition with the sense of the current reversed by means of the switch S. The quantities x' and x'' are pure numbers and ν is measured in terms of revolutions per second. The mutual inductance can be calculated as indicated in the preceding chapter and is merely a geometrical factor times the constant μ_0. Thus R is determined in terms of this arbitrarily chosen quantity and the units of length and time. With the necessary precautions and refinements this method is used in the national standardizing laboratories for determining the values of resistance standards to an accuracy of about 2 parts in 10^5.

The constant κ_0 was included in Coulomb's law of force between electric charges in anticipation of these later electromagnetic results. It was so chosen that if the charges are expressed in coulombs and their separation in meters, the force between them is given in newtons. In that sense the equation served as a temporary definition of charge though the actual definition of the coulomb as the ampere-second had to be postponed till the ampere was defined through the current balance. Once a standard of current has been obtained, it is theoretically possible to place known quantities of charge a certain distance apart and measure the force they exert on one another, but it would be very difficult to achieve the desired accuracy in such an experiment. A much more satisfactory method is to compare a capacity C, which is equal to κ_0 times a geometrical factor, with the standard of resistance which has just been determined.

FIG. 10.6.—Method of comparing a capacity and a resistance and hence of measuring the constant κ_0.

This can be accomplished by means of the circuit shown in Fig. 10.6 which is essentially a Wheatstone bridge. The vibrator v is driven at a constant rate of ν times a second between the contacts indicated by arrows. Thus the condenser C is alternately charged to some potential V and discharged ν times a second, and a quantity of charge equal to νCV is transferred from A to B per second. This is equivalent to an average current of νCV or the effective resistance of the arm is $V/i = 1/\nu C$. Hence, if the inertia of the galvanometer is such that it only responds to the average current through it, the balance condition is

$$\frac{R_1}{R_3} = R_2 \nu C$$

Since $C = \kappa_0 C'$, where C' is a geometrical factor (the area over the separation for a parallel-plate condenser)

$$\kappa_0 = \frac{R_1}{R_2R_3\nu C'}$$

In this equation all the quantities have been strictly defined and since the method is capable of high precision, κ_0 can be accurately measured. Using essentially this method and employing many refinements to ensure the highest possible accuracy, Rosa and Dorsey determined κ_0 to be 8.851×10^{-12} farad per meter with an accuracy of about 4 parts in 10^5, which is approximately the accuracy in the determination of the standard of resistance. This is a very important experimental physical constant since it is involved in all quantitative electrostatic calculations. It also enters electrical-circuit theory through the use of capacities and will be seen to play an important role in the discussion of radiation.

10.5. Electromagnetic Instruments for Measuring Current, Charge, and Flux.—Electromagnetic instruments for which the deflection of the moving element depends on the magnetic interaction between a current-carrying coil and a permanent magnet are known in general as galvanometers. They are of basic importance in all branches of electrical measurement, and a discussion of their characteristics will serve as an excellent illustration of the ideas contained in this chapter and the preceding one. The general nature of permanent magnets will not be taken up until the following chapter, but some familiarity with them is assumed at this point. They are specially prepared pieces of hard steel which possess a permanent magnetic moment. In this respect they are analogous to superconductors in which circulating currents have been established. The magnetic field which a permanent magnet produces in its neighborhood is very similar to that which would be produced by a solenoid closely wound over the surface of the magnet. These electromagnetic instruments are divided into two classes for convenience: (*a*) the moving-magnet type, in which a small permanent magnet rotates in the magnetic field produced by a current flowing in a stationary coil; and (*b*) the moving-coil or D'Arsonval type, in which a current-carrying coil rotates in the field of a permanent magnet.

Moving-magnet galvanometers are of two types, depending on the nature of the restoring torque acting on the magnet. The tangent galvanometer illustrated in Fig. 10.7 represents a type of instrument in which the restoring torque on the magnetic needle is due to the earth's magnetic field. The magnetic needle is mounted on a jeweled pivot so that it is free to rotate in a horizontal plane. If $\mathbf{m}$ is the moment of the needle and $\mathbf{H}_e$ is the earth's magnetic field, the torque exerted by it on the needle is $\mathbf{m} \times \mathbf{H}_e$ and the equilibrium position of the needle is parallel to the earth's field. The magnetic needle is at the center of a coil of wire in a vertical plane parallel to the earth's field. The radius

of this coil is large in comparison with the linear dimensions of the magnet. If the coil contains n turns and has an effective radius b in meters, the magnetic field at the center due to a current i in the coil is given by Eq. (9.11) as $H_c = ni/2b$. In equilibrium the torques exerted on the

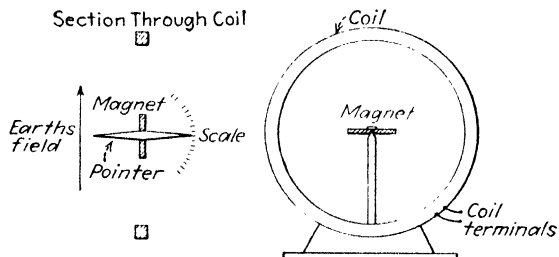



FIG. 10.7.—The tangent galvanometer.

magnet by the current and the earth's field exactly balance one another or, if θ is the deflection of the needle away from the earth's field

$$mH_e \sin \theta = mH_c \cos \theta$$

or

$$\tan \theta = \frac{n}{2bH_e} i$$

Thus the current is proportional to the tangent of the angle of displacement. It will be seen in the following chapter that it is possible to make an absolute measurement of H_e and as the other quantities in the coefficient of i can be directly measured, the tangent galvanometer can be used to make an absolute measurement of current. However, such a measurement is less accurate than can be made with a current balance. Unless the magnet is very small in comparison with the coil, the field acting on the magnet cannot be considered uniform. An alternative arrangement is to place the magnet in the central region between two coaxial coils separated by a distance equal to their radius (Helmholtz coils), for the field is more uniform and simply calculable in this region.

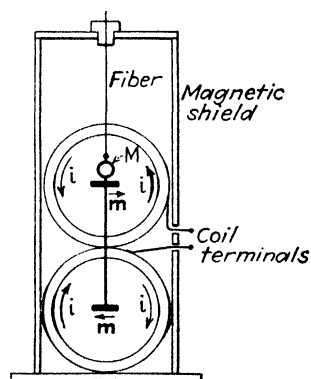



FIG. 10.8.—Astatic galvanometer.

A type of moving-magnet instrument in which the restoring torque is due to a suspending fiber is known as an astatic galvanometer. This is illustrated in Fig. 10.8. A light rigid vertical rod carries two small coplanar permanent magnets of equal moment which are affixed perpendicularly to it with their moments oppositely directed. Thus a

uniform field such as that of the earth exerts no net torque upon the system. To reduce the effect of stray nonuniform fields the instrument is surrounded by magnetic shields of soft iron. The moving system carries a mirror by means of which the deflection of a reflected beam of light can be observed. It is suspended from its upper end by a fine quartz fiber in such a way that each magnet is at the center of a coil or pair of coils. The current is sent through the upper and lower coils in opposite directions so that the torques on the two magnets are in the same sense. If m is the moment of each magnet and n the number of turns surrounding it, the torque due to a current i is $(mni \cos \theta)/b$, where b is the radius of the coils and θ the angular deflection of the magnets from the plane of the coils. This is opposed by the restoring

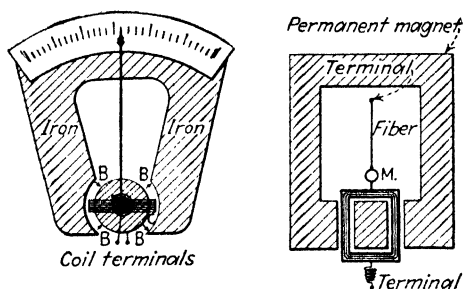



FIG. 10.9.—Pointer and mirror galvanometers.

torque of the fiber which for small displacements can be taken as proportional to the deflection, *i.e.*, it can be written $k\theta$, where k is the restoring torque per unit deflection. Since θ is assumed small, $\cos \theta$ may be taken as unity and

$$i = \frac{kb}{mn}\theta$$

Since k and b can be made small and m and n large, it is possible to construct an instrument of this type that has a high sensitivity, *i.e.*, a large ratio of θ to i. In fact, instruments of this type with the necessary refinements can be made more sensitive than the moving-coil instruments which will be discussed later; currents of the order of 10^{-12} amp. can be readily detected. However, they are more affected by stray magnetic fields and the damping is more difficult to control so they are less generally useful than moving-coil instruments. As it is not practicable to make an absolute determination of m, the instrument is calibrated by passing a known current through the coils and the coefficient of θ or the ratio of the current to the displacement of a beam of light reflected from the mirror M is determined experimentally.

The moving coil instrument is the basic element in most commonly encountered galvanometers, ammeters, and voltmeters. It consists

of a coil of wire which is free to rotate in the field of a permanent magnet. The center of the coil is generally occupied by a cylindrical iron block known as the armature, and the magnet is shaped in such a way that its ends or poles form an annular space about the armature in which the coil can rotate. This construction results in a large and uniform radial induction in the region occupied by the sides of the coil. In the more rugged type of instrument, which is illustrated at the left in Fig. 10.9, the coil is mounted in jeweled bearings and carries a pointer, by means of which its deflection can be observed. A spiral spring near one bearing provides the restoring torque and the current enters and leaves the coil through this spring and a light flexible wire near the other bearing. The limit of sensitivity of such an instrument is approximately 10^{-6} amp. per division. That is, the *galvanometer constant* which is defined as the ratio of the current to the displacement and written S is equal to approximately 10^{-6} for this type of instrument.

In the less rugged but more sensitive type illustrated at the right in Fig. 10.9 the coil is suspended by means of a fine phosphor-bronze fiber which exerts a smaller restoring torque than the spiral spring used in the pivot type of instrument. The current enters and leaves the coil by means of this fiber and a loose helix of fine wire which exerts a negligible restoring torque. The deflection is observed by viewing a scale in the mirror carried by the coil or by the deflection of a beam of light incident on this mirror. Since the restoring torque of the fiber is small, the sensitivity of this type of instrument can be made very large. A very sensitive galvanometer of this type will give a deflection of 1 mm. on a scale 1 m. away for a current of about 10^{-11} amp., *i.e.*, $S = 10^{-11}$. It will be seen later that a galvanometer of small internal resistance is most efficient for detecting small emfs. in low-resistance circuits. In this service instruments are classed according to their voltage sensitivity, *i.e.*, the fraction of a volt which when applied to their terminals produces a unit deflection. Instruments are available which will produce 1 mm. deflection on a scale distant 1 m. for an emf. of 10^{-9} volt. This is approximately 10^5 times the voltage sensitivity of a sensitive electrometer. A galvanometer is the most sensitive instrument for measuring emfs. developed in low-resistance circuits. The considerations underlying the choice of a galvanometer for a particular purpose will be brought out in the subsequent analysis.[1]

Frequently it is desirable to reduce the sensitivity of the instrument

[1] Galvanometer sensitivities are sometimes quoted in megohms. The megohm sensitivity is the number of megohms that must be placed in series with the instrument and an emf. of 1 volt in order to limit the deflection to one division of the scale. Thus, since the galvanometer resistance is small, the megohm sensitivity is numerically equal to the sensitivity in divisions per microampere.

by a known amount. This can be accomplished and at the same time the proper damping resistance maintained in series with the instrument by means of an Ayrton shunt (Sec. 4.5). Also, the ordinary ammeters and voltmeters are merely pivot galvanometers which have had their sensitivities reduced in the proper manner. If a resistance R is placed in shunt across the terminals of a galvanometer and a current i flows to the shunted combination, the current through the galvanometer is given by $i_g = Ri/(R + R_g)$, where R_g is the resistance of the galvanometer coil. Since i_g is equal to the product of the deflection d and the galvanometer constant

$$\frac{i}{d} = \frac{S(R + R_g)}{R}$$

This gives the ratio of the number of amperes in the external circuit to the deflection registered by the instrument. By the proper choice of the shunt resistance R any desired scale factor can be obtained. Similarly, if a resistance R is placed in series with the galvanometer coil, the deflection of the instrument can be taken as a measure of the potential difference applied across the combination. Since $V = i_g(R + R_g)$ and $i_g = dS$

$$\frac{V}{d} = S(R + R_g)$$

Thus the number of volts corresponding to a unit deflection of the instrument can be determined by a suitable choice of the external resistance R. In practice the shunt or series resistance associated with the instrument is not calculated from the galvanometer constants but is adjusted empirically for the proper deflection at a known current or potential difference as measured by a potentiometer circuit.

To analyze the static characteristic of the moving-coil galvanometer assume that the coil has n turns and is rectangular in shape. If the breadth of the coil is a and its height b, the torque exerted upon it by the induction B of the permanent magnet is seen to be $nabBi$, where i is the current. This is opposed by the restoring torque of the spring or fiber which can be written $k\theta$. Thus, in equilibrium

$$i = \frac{k}{nabB}\theta$$

or the galvanometer constant is given by $S = k/nabB$. For the maximum sensitivity, B should be as large as possible. The restoring constant k cannot be decreased beyond a certain limit and it is not feasible to make a or b greater than a few centimeters. Nor can n be increased indefinitely since the product of n and the cross section of the wire are limited by the space available in the annular region between the magnet

poles. Assuming that g is the cross section of the wire and g' the maximum cross section allowable for the coil, the optimum condition is $ng = g'$. The resistance of the galvanometer coil is also determined by n and g, for if ρ is the specific resistance of the copper in the wire and the space occupied by the insulation is neglected,

$$R_g = \frac{2\rho n(a + b)}{g} = \frac{2\rho n^2(a + b)}{g'}$$

Eliminating n between this expression and that for the sensitivity of the instrument

$$\theta = \frac{Bab}{k}\sqrt{\frac{g'}{2\rho(a + b)}}R_g^{1/2}i$$

Thus the deflection is proportional to $R_g^{1/2}i$ and will be a maximum for a maximum value of $R_g i^2$. This is the same relation as that for the maximum dissipation of power in a resistance R_g and the galvanometer can be considered as essentially a power device. If an emf. $\mathcal{E}$ is generated in a circuit of resistance R in series with the galvanometer, the deflection of the instrument is proportional to $R_g^{1/2}/(R + R_g)$. If R can be varied, the largest deflection will be obtained if it is made as small as possible. However, if the galvanometer can be chosen, one should be selected for which R_g is as nearly as possible equal to R, for $R_g = R$ is the condition for the maximum of this fraction. Thus the resistance of a galvanometer should be matched to that of the external circuit in which it is to be used. For instance, a low-resistance instrument should be used to measure the potential developed by a thermocouple or in conjunction with a low-resistance potentiometer whereas a high-resistance galvanometer should be used to measure the potential of a large thermopile or of a high-resistance cell, or with a high-resistance potentiometer. Of course, in those comparatively rare instances in which the current rather than the emf. of the circuit is constant a galvanometer with a high current sensitivity, which implies a high resistance, should be used.

For convenience in operation the movement should be quite highly damped so that it achieves its equilibrium deflection in as short a time as possible. Part of this damping is produced by the movement of the needle or coil through the air and part is contributed by the electromagnetic reaction of the circuit. These two terms are of the same order of magnitude for a moving-magnet galvanometer, but the second is generally much greater for the moving-coil instrument. Consider first the case of the moving-magnet type and assume that the magnet has a moment m which may be thought of as the product of μ_0 and a circulating current i' enclosing an area a. The flux through this area per unit current in the coil is $(\mu_0 na/2b) \sin \theta$ if θ is the angle between the moment of the

magnet and the plane of the coil. Since the flux through the coil per unit i' of the magnet must also be equal to this quantity and as $m = \mu_0 i'a$, the flux through the coil due to the magnet is $(mn/2b) \sin \theta$. The emf. induced in the coil by a changing flux is $-\frac{\partial\phi}{\partial t} = -\frac{\partial\phi}{\partial\theta}\frac{d\theta}{dt}$, and this gives rise to a current $-\frac{1}{R}\frac{\partial\phi}{\partial\theta}\frac{d\theta}{dt}$, where R is the total resistance of the coil and any external circuit. The retarding torque is the product of this current and $\partial\phi/\partial\theta$, or since totals can be written as well as partial derivatives

$$T_r = -\frac{1}{R}\left(\frac{d\phi}{d\theta}\right)^2\frac{d\theta}{dt}$$

Hence the electromagnetic retarding torque due to the two coils of an astatic galvanometer is

$$T_r = -\frac{1}{R}\left(\frac{mn \cos \theta}{b}\right)^2\frac{d\theta}{dt}$$

For the moving-coil galvanometer, since $d\phi/d\theta = nabB$

$$T_r = -\frac{1}{R}(nabB)^2\frac{d\theta}{dt}$$

Limiting the discussion to the case in which θ is small enough so that the cosine is approximately unity, it is seen that the terms in brackets are equal to the ratio of the restoring constant to the galvanometer constant (the restoring constant would be equal to mH_e for a tangent galvanometer). In general for any of these galvanometers

$$T_r = -\frac{1}{R}\left(\frac{k}{S}\right)^2\frac{d\theta}{dt}$$

Neglecting air damping, the equation of motion of a galvanometer is

$$I\frac{d^2\theta}{dt^2} = -\frac{1}{R}\left(\frac{k}{S}\right)^2\frac{d\theta}{dt} - k\theta + \frac{ki}{S} \tag{10.9}$$

where I is the moment of inertia of the moving element. The discussion will be limited to the case of a constant current for which the equilibrium deflection is i/S. If the deflection θ is measured from the equilibrium position, the equation does not contain the final term and in the notation of Appendix C would be written

$$\frac{d^2\theta}{dt^2} + 2a\frac{d\theta}{dt} + \frac{k}{I}\theta = 0$$

where $a = k^2/2RIS^2$, which is of the form of Eq. (C.2). The periodic solution, which is the one of principal interest, is given by Eq. (C.13). The particular form that this solution takes depends on the initial

displacement and rate of displacement. An instance which represents the general nature of the motion and which will be found applicable to the ballistic use of the instrument is obtained by assuming that the galvanometer is given an impulse when in its equilibrium position, *i.e.*, $\theta_0 = 0$, where θ_0 is the displacement from the equilibrium position at $t = 0$. In this case the constants of Eq. (C.13) become $\delta' = 0$ and $Q = \omega_0$, where ω_0 is the initial angular velocity and the equation giving the displacement at any time becomes

$$\theta = \frac{\omega_0}{m'} e^{-at} \sin m't \tag{10.10}$$

where m' is written for $\left(\frac{k}{I} - a^2\right)^{1/2}$. Thus the galvanometer movement

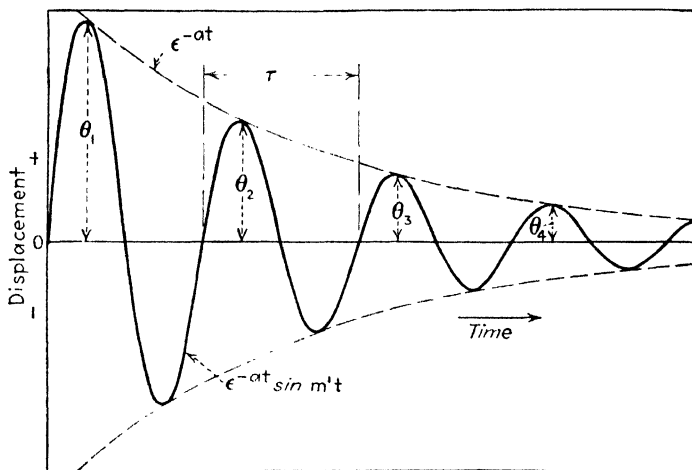



FIG. 10.10.—Damped simple harmonic vibration.

executes the type of damped simple harmonic motion illustrated in Fig. 10.10. The zeros of the function are evidently periodic with a period $\tau = 2\pi/m'$. The maxima are obtained by setting $d\theta/dt = 0$, which yields

$$\tan m't = \frac{m'}{a}$$

hence these are also periodic with the same period. The ratios of the heights of successive maxima are evidently the ratios of values of θ separated in time by τ, or since the periodic factor returns to the same value after this interval

$$\frac{\theta_2}{\theta_1} = \frac{\theta_3}{\theta_2} = \frac{\theta_4}{\theta_3} = \cdots = e^{-a\tau}. = e^{-\delta}$$

where δ is known as the *logarithmic decrement.*

If the damping is small, the period is approximately equal to the free period, $\tau_f = 2\pi(I/k)^{1/2}$. In this case $\delta = a\tau_f$ or, since a is inversely proportional to R, the decrement is also inversely proportional to the resistance of the circuit. In the ordinary use of a galvanometer it is most convenient to have the instrument approximately critically damped, for then it takes up its equilibrium deflection in the shortest time (Appendix C). This condition corresponds to $m' = 0$ or $a_c^2 = k/I$ which can be written $a_c = 2\pi/\tau_f$. If δ_1 is a small decrement corresponding to a large circuit resistance R_1, τ_f can be eliminated between the expressions for δ_1 and a_c to yield

$$a_1 = \frac{\delta_1 a_c}{2\pi}$$

or since a is inversely proportional to the circuit resistance

$$R_c = \frac{\delta_1 R_1}{2\pi}$$

where R_c is the circuit resistance corresponding to critical damping. This expression enables one to determine the circuit resistance for critical damping in terms of the decrement for a large known circuit resistance. As the resistance of the galvanometer is small in comparison with R_c, the equation can be interpreted approximately as giving the external resistance necessary for critical damping. The function of an Ayrton shunt, for instance, is to vary the sensitivity of the galvanometer and at the same time maintain the resistance of the external circuit at approximately this critical value.

An important application of the galvanometer is in the so-called "ballistic" service for the measurement of charge. Any galvanometer can be used in this way, but special instruments having a large moment of inertia are designed for this purpose. The requirement is that the galvanometer movement should not deflect appreciably in a brief interval δt during which a current i passes through the instrument. During that time it acquires an impulse which causes it to swing away from its equilibrium position. Multiply both sides of Eq. (10.9) by dt and integrate the terms over the short interval δt. The term on the left becomes $I(d\theta/dt)_0^{\delta t}$ and since θ is equal to zero at both the beginning and end of the interval, the first two terms on the right vanish. The third term on the right is equal to $\frac{ki}{S}\delta t$ or since $i\,\delta t$ is equal to the total charge q that traverses the instrument, the term can be written kq/S. The movement has no angular velocity at the beginning of the interval so the integral of Eq. (10.9) over δt becomes

$$I\omega_0 = \frac{kq}{S}$$

where ω_0 is written for $(d\theta/dt)_{\delta t}$ which is the angular velocity imparted to the movement during the interval that causes it to swing away from its equilibrium position. Inserting this value of ω_0 in Eq. (10.10), the subsequent deflection at any time is found to be

$$\theta = \frac{k}{SIm'} q e^{-at} \sin m't$$

Thus the subsequent deflection of the instrument is proportional to the charge that passed through it during the initial period. The charge on a condenser can be measured, for instance, by connecting it across the terminals of a galvanometer and observing the deflection at any later time if the constants of the instrument are known. It is most convenient, however, to use the first maximum as a measure of the charge. From the previous discussion this is seen to occur at a time $t' = 1/m' \tan^{-1} (m'/a)$ and as the sine function at this time is equal to $m'/(m'^2 + a^2)^{1/2}$, this first maximum deflection is given by

$$\theta_m = \frac{2\pi}{S\tau_f} q e^{-\left(\frac{\delta}{2\pi}\right)\tan^{-1}\frac{2\pi}{\delta}}$$

since $a/m' = \delta/2\pi$. To obtain a large deflection and for ease of calculation the damping of the instrument should be as small as possible. For this reason moving-magnet instruments are particularly suitable and the moving-coil type should be used if possible on open circuit. This is readily accomplished in measuring the charge of a condenser since the condenser effectively produces an open circuit for oscillations as slow as those of a galvanometer. If δ is very small, $\tan^{-1} (2\pi/\delta)$ is approximately equal to $\pi/2$ or the exponent is $-\frac{\delta}{4}$ and

$$\theta_m = \frac{2\pi}{S\tau_f}\left(1 - \frac{\delta}{4}\right)q \tag{10.11}$$

This is the most convenient expression to use for the measurement of charge in terms of a ballistic-galvanometer deflection. In place of measuring the galvanometer constants the instrument can be calibrated by charging a known capacity C to a known potential V and noting the deflection produced when the condenser is discharged through the instrument. This determines the coefficient of q as $\theta'/(VC)$ where θ' is the observed deflection.

It is also possible to measure flux with a ballistic galvanometer through the use of Eqs. (10.1) and (10.2). If a flux ϕ linking a circuit in series with a galvanometer is changing at the rate $d\phi/dt$ the differential equation of the circuit is

$$L\frac{di}{dt} + Ri = -\frac{d\phi}{dt}$$

If this equation is multiplied by dt and integrated over the time interval δt, the first term on the left vanishes since the current is zero at both the beginning and end of the interval, the second term on the left is equal to Rq, where q is the circulated charge, and the term on the right is equal to the difference between the flux linked by the circuit at the beginning and end of the interval. Therefore

FIG. 10.11.—Calibration of a ballistic galvanometer with a standard solenoid.

$$q = \frac{\delta\phi}{R}$$

where $\delta\phi$ is the change in flux linkage. Thus a flux measurement can be reduced to the measurement of a charge. In terms of Eq. (10.11)

$$\theta_m = \frac{2\pi}{SR\tau_f}\left(1 - \frac{\delta}{4}\right)\delta\phi \qquad (10.12)$$

This equation is applicable only if the damping is small; hence, if the resistance of the galvanometer circuit is small, the circuit must be opened immediately after the current impulse has traversed it. If the change in flux is induced by the breaking of a circuit, the galvanometer circuit itself must be broken afterward. Special keys are available for this purpose. This procedure permits the determination of the ballistic constant of a galvanometer by the use of a standard solenoid. The key K of Fig. 10.11 first breaks the circuit through the solenoid and then breaks the circuit composed of the galvanometer and a coil of m turns wound over the solenoid near its center. If the solenoid of cross-sectional area A and n' turns per unit length was carrying a current i, the flux linkage through this coil was $\mu_0 mAn'i$ before the circuit was broken. If the observed deflection is θ'_m

$$\theta'_m = Q'\mu_0 mn'Ai$$

where Q' is the coefficient of $\delta\phi$ in Eq. (10.12). Since Q' is the only quantity that is not known, it can be determined through this equation and the galvanometer is calibrated for the particular circuit resistance R. If it is to be used for flux measurement with a different coil, the resistance of the circuit must be kept the same, which is generally accomplished by including the coil in the galvanometer circuit during calibration. If this is not done, the resistance of the galvanometer circuit must be determined and the necessary correction applied for any alteration in it.

The magnetic induction in any small region can be determined by means of a search coil and ballistic galvanometer. The search coil is

simply a small plane coil with flexible leads as shown at the left in Fig. 10.12. If it has an area A and contains n turns, the galvanometer throw (first maximum deflection) will be $nAQ'B$ if it is inserted or withdrawn from a region where the normal induction through it is B. However, it is necessary to break the circuit after insertion or withdrawal to obtain a free throw and this is frequently inconvenient. Hence another type of instrument known as a *fluxmeter* is used for this purpose. A fluxmeter is simply a moving-coil galvanometer with its coil mounted in such a way that the suspending fiber exerts a negligible restoring torque. It will rest equally well at any position within the limits of its motion. Equation (10.9) describes the motion of this type of instrument as well

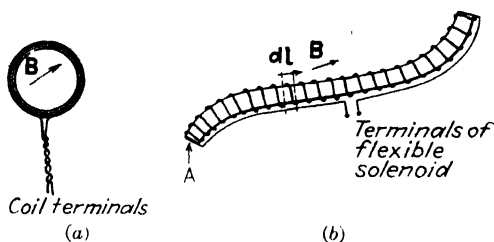



FIG. 10.12.—(a) Search coil. (b) Magnetic potentiometer.

and though both k and S are small their ratio is equal to $nabB$ which will be written C for brevity. Thus the equation is

$$I\frac{d^2\theta}{dt^2} + \frac{C^2}{R}\frac{d\theta}{dt} = Ci$$

This equation can be integrated immediately and since $d\theta/dt$ is zero at the beginning and end of the motion and $\int i\,dt = q$, the change in deflection $\delta\theta$ is given by

$$\delta\theta = \frac{Rq}{C}$$

For a flux measurement $q = \delta\phi/R$ or

$$\delta\theta = \frac{\delta\phi}{C}$$

The constant C is determined by calibration with a standard solenoid and the scale of the instrument is marked off in divisions representing webers or maxwells (1 weber = 10^8 maxwells). Since the restoring torque is very small, the time that elapses during the motion of the search coil is of little importance.

In making magnetic measurements it is frequently necessary to recalibrate the galvanometer. This is the case, for instance, if the resistance in the circuit is changed from time to time. The Hibbert magnetic standard is a convenient instrument for this purpose. A

section through this device is shown in Fig. 10.13. A cylindrically symmetrical permanent magnet is so shaped that there is an open annular ring in its upper surface. In this opening there is a strong and constant induction parallel to the surface. A brass cylinder which is of the proper diameter to drop through this annular opening carries a coil of fine wire wound upon its surface. The terminals of the coil are brought out to binding posts on the insulating disk that closes the top of the cylinder. When the cylinder is allowed to drop through the annular opening, the flux linkage of the coil changes by an amount characteristic of the particular instrument. It is not an absolute device, but must originally be calibrated against a standard solenoid. However, it is very convenient and rugged and will retain its calibration over considerable periods of time if the permanent magnet is of good quality and carefully handled. If the flux through the gap is ϕ' and there are n' turns on the coil, the charge circulated through the galvanometer when the cylinder falls through the opening is $n'\phi'/R$, where R is the resistance of the circuit. Let this correspond to a deflection θ'. If a search coil of n turns and area A is also in series with the instrument and its withdrawal from a region yields a deflection θ, the normal induction in the region is given by $n'\phi'\theta/nA\theta'$. The quantity $n'\phi'$ is supplied by the maker or obtained by calibration.

FIG. 10.13.—Hibbert magnetic standard.

A useful device for magnetic measurements in conjunction with a ballistic galvanometer or fluxmeter is the magnetic potentiometer shown at the right in Fig. 10.12. It is simply a long flexible solenoid which is used in the same way as a search coil. It is generally made by wrapping several thousand turns of fine wire upon a leather strap and suitably serving it to protect the winding. The ends of the wire are brought out on a terminal block. If the device is placed in a magnetic field, the total flux linking it is the sum of the product of the area of cross section and the normal component of the induction at that point for each turn. This can be written

$$\phi = n'A\int_l \mathbf{B} \cdot d\mathbf{l} = \mu_0 n'A\mathcal{H}$$

where n' is the number of turns per unit length and A is the area of cross section of the winding. $\mathcal{H}$ is the magnetomotive force between the ends of the coil. Thus the flux linkage is a measure of the magnetomotive force existing between the ends of the solenoid. The constant of the instrument, $n'A$, is supplied by the maker or can be obtained by calibration. The calibration is performed by linking the flexible solenoid

with a few turns of wire, say m, and passing a current i through these turns. The mmf. established is mi; hence $n'A = \phi/\mu_0 mi$, and ϕ is given by the scale of a fluxmeter. The magnetic potentiometer is particularly useful for measuring the leakage flux outside large magnets. This is determined by measuring the magnetomotive force between points on its surface. It may also be used to determine the number of turns in an air-cored coil. The instrument is threaded through the coil and its ends brought together. If a current i is sent through the coil, the mmf. established is mi, where m is the unknown number of turns in the coil. This measurement is essentially the reverse of the calibration procedure.

10.6. Magnetic Induction Accelerators.—The electromotive force induced by changing flux can also be used either alone or in conjunction

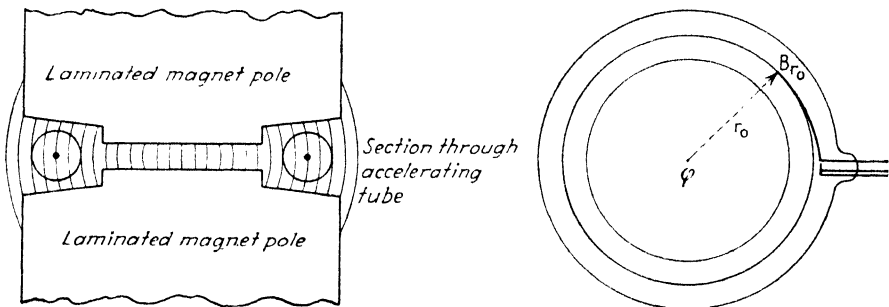



FIG. 10.14.—Schematic representation of a betatron.

with electric fields to accelerate charged particles to very high velocities. The betatron[1] was the first device of this type to be operated successfully. This machine consists of a large alternating-current magnet that produces a field across an air gap having radial symmetry about the axis of the poles as shown schematically in Fig. 10.14. The region in which the particles are accelerated is an evacuated toroidal tube which is coaxial with the pole pieces and lies near the periphery of the field. The alternating field is of the form $H_m \sin \omega t$; and if electrons are to be accelerated, they are injected circumferentially by means of a filament and pierced anode near the center of the toroid section shortly after the field has passed through its zero value in the proper sense to bend the electron beam around within the toroid. The first-order equation for the equilibrium orbit can be derived from Eqs. (9.5) and (10.1). Assuming that the electrons move in a circle of constant radius r_0, the only component of momentum is the tangential one, and this may be written by Eq. (9.5) as

$$p_t = mu_t = er_0 B_{r_0}$$

[1] KERST, *Phys. Rev.*, **58**, 841 (1940); **59**, 110 (1941); KERST and SERBER, *Phys. Rev.*, **60**, 53 (1941).

The tangential force by Eq. (10.1) is equal to the rate of change of mometum or

$$\frac{dp_t}{dt} = \frac{e}{2\pi r_0}\frac{d\varphi}{dt}$$

where φ is the total flux through the circle of radius r_0. If $p_t = 0$ for $\varphi = 0$, which is the approximate condition for the starting of the electrons in their orbit,

$$p_t = \frac{e}{2\pi r_0}\varphi$$

Eliminating p_t the necessary condition between φ and B_{r_0} is seen to be

$$B_{r_0} = \frac{1}{2}\frac{\varphi}{\pi r_0^2}$$

Thus if the field at r_0 is one-half the average field over the orbit, the electrons will be continuously accelerated during the quarter cycle from $B = 0$ to $B = B_{\max}$, remaining always in the circle of radius r_0. If the ultimate velocity is much less than that of light, the maximum kinetic energy can be written as $(1/2m)p^2{}_{\max}$, which in electron volts is $\frac{1}{2}(e/m)r_0^2B_m^2$, where B_m is written for the maximum value of B at r_0. In the case of electrons the velocity becomes comparable to that of light when the energy becomes high, and therefore the relativistic expression mc^2 for the energy must be used. Since $p^2 = (m^2 - m_0^2)c^2$ the kinetic energy is

$$(KE)_{\max} = c(p^2{}_{\max} + m_0^2c^2)^{1/2} - m_0c^2$$

and $p_{\max}$ is, of course, given by er_0B_m. Putting in numerical values for c and e/m it is seen that m_0c^2 corresponds to an energy of about half a million electron volts. If r_0B_m is so large that $p_{\max} >> m_0c$. The kinetic energy becomes approximately er_0B_mc, or in electron volts r_0B_mc. It is not difficult to achieve dimensions and fields that justify this approximation. Taking B_m as 3,000 gauss or 0.3 weber per square meter and r_0 as 0.25 m., the energy in electron volts is seen to be 22.5 million electron volts.

The actual design and operation of the betatron is, of course, more involved than the simple discussion given above would indicate. The design and fabrication of the pole pieces of the magnet are quite critical. The magnet must be laminated radially to reduce eddy current losses, and the radial contour must be such as to fulfill both the equilibrium orbit conditions and conditions of stability (Sec. 9.2). Automatic circuits must be employed to inject the electrons at the proper time in the alternating current cycle, and auxiliary coils must be used to expand or contract the orbit at the proper time to make the beam of electrons strike a target so that they will not be decelerated during the subsequent quarter

cycle. By means of such coils and local magnetic shielding the electron beam may be brought out of the toroid through a thin window properly placed. The radiation of energy by the circulating electrons imposes an upper limit of a few hundred million electron volts in straight betatron operation. However, by employing a combination of induction acceleration and electric acceleration, which can be brought about by inserting electrodes within the toroid, this limitation can be overcome. The machine employing both principles of operation is the synchrotron[1] referred to in Sec. 9.2, and it offers the most promising means for achieving electron or ion energies in the billion-electron-volt region.

10.7. Magnetic Characteristics of Atomic Systems.—A consideration of the magnetic properties of matter in bulk will be postponed until the following chapter, but an elementary discussion of the magnetic properties of atomic systems will be given here as an introduction to the larger subject and an illustration of general magnetic principles. We know that an atom of a substance is composed of a relatively massive nucleus with a characteristic positive charge surrounded by a sufficient number of electrons to render the system electrically neutral. This system is in equilibrium under forces of three types: (*a*) electric, (*b*) magnetic, (*c*) nonclassical forces for a description of which quantum mechanics must be invoked. The forces that associate atoms together in molecules and larger aggregates such as liquids and crystals are also of these three general types. In this section we are interested chiefly in the forces of type (*b*) arising from the magnetic moments associated with atoms, their components, and their aggregates, although the forces of types (*a*) and (*c*) are also important in their effect on the establishment of equilibrium conditions. The situation even in regard to magnetic interaction alone is inherently very complex. The atomic nucleus possesses an intrinsic magnetic moment and angular momentum as does each of the electrons as well. The motions of the electrons in the atom produce magnetic moments of the amperian current vortex type, and the coupling of atoms in molecules and interaction of neighboring atoms in crystals affect the orientations of all of these magnetic moments. However, except in the cases of ferromagnetic substances the forces of interaction between neighboring atoms are not such as to produce a net magnetic moment in a macroscopic volume; and hence as far as the magnetic moment is concerned, it may be considered as arising from individual atoms or molecules that are randomly oriented unless acted upon by some externally applied field. The strength of the interaction between the nuclear and electronic components of an individual atom or molecule, however, is such that a mean directional relation between

[1] BOHM and FOLDY, *Phys. Rev.*, **72**, 649 (1947).

these components is maintained, and these systems do react as a whole to weak applied fields. If the applied field is increased, the energy of interaction between it and these components either individually or as subgroups increases to such an extent that the intra-atomic coupling is broken down, and eventually in the case of very large external fields the nuclei and electrons composing an atom have their motions oriented almost entirely by the external field, and the intra-atomic interactions become almost negligible.

An atom or molecule may thus exhibit quite different phenomena at different external field strengths. However, the existence of a characteristic angular momentum and magnetic moment associated with nuclei, electrons, and electronic orbital motion brings about a marked similarity in the interaction of each of these or the aggregate they compose with a magnetic field. The atomic components behave like magnetic spinning tops, and the atom resembles a complex of such tops, which, however, reacts much as a single top at low fields. The energy of magnetic interaction between the nucleus and the external electrons is quite small, so very moderate external fields decouple the nuclear orientation from that of the electron structure, and stronger fields are required to bring about a decoupling of the intrinsic and orbital magnetic moments of the various electrons forming the external atomic structure. A study of the interaction of atoms with magnetic fields elucidates the nature and magnitude of interatomic magnetic forces and also yields the ratios of magnetic moment to angular momentum for the atom as a whole and for each of its components. This ratio for an atom as a whole is a function of the applied field because of the variation of coupling of its components, but the ratio of intrinsic moment to angular momentum or spin for a nucleus or an electron and this ratio for a single circulating charge are invariants. As an illustration the orbital motion of an electron will be considered.

When a top with a constant angular momentum is acted on by a torque which tends to change the direction of its axis, it is observed to acquire a rotation or to precess about a third axis. The vector $\mathbf{p}$ of Fig. 10.15 represents the angular momentum of the top and $\mathbf{dp}$ its change in a time dt. The magnitude of $\mathbf{dp}$ is $p\,d\theta$ and it is perpendicular to $\mathbf{p}$ and to $\mathbf{d\theta}$. The induced angular velocity of precession is $\mathbf{d\theta}/dt$, which will be written $\boldsymbol{\omega}$. Hence

$$\mathbf{dp} = \mathbf{d\theta} \times \mathbf{p} = \boldsymbol{\omega} \times \mathbf{p}\,dt$$

and since by definition the torque $\mathbf{T}$ is equal to the rate of change of angular momentum

$$\mathbf{T} = \frac{\mathbf{dp}}{dt} = \boldsymbol{\omega} \times \mathbf{p} \tag{10.13}$$

Now, if the magnetic moment associated with the atom or atomic component is $\mathbf{m}$ the torque exerted on it by a field $\mathbf{H}$ is $\mathbf{m} \times \mathbf{H}$ and the precession induced by a field $\mathbf{H}$ is given by

$$\mathbf{m} \times \mathbf{H} = \boldsymbol{\omega} \times \mathbf{p}$$

$\mathbf{m}$ and $\mathbf{p}$ will be taken as parallel or antiparallel to one another in the simplest case and hence

$$\boldsymbol{\omega} = -\frac{\mathrm{p}}{\mathrm{m}}\mathbf{H} \tag{10.14}$$

Thus the induced precession, which is known as a *Larmor precession,* is parallel to $\mathbf{H}$ and proportional to it as well as proportional to the ratio of the magnetic and mechanical moments.

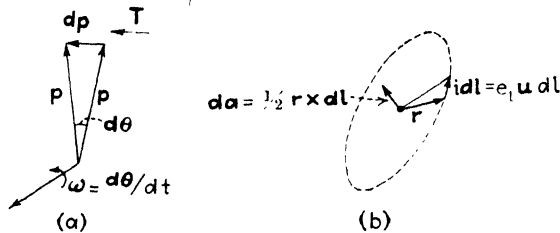



FIG. 10.15.—(*a*) Vector analysis of precession. (*b*) Circulation of an electronic current about a nucleus.

Both $\mathbf{m}$ and $\mathbf{p}$ are readily calculable for a circulating electron. By definition $\mathbf{m} = \mu_0 i\mathbf{a}$ and from Fig. 10.15 $\mathbf{a} = \int \mathbf{da} = \frac{1}{2}\int \mathbf{r} \times \mathbf{dl}$, so

$$\mathbf{m} = \tfrac{1}{2}\mu_0 i\int \mathbf{r} \times \mathbf{dl} = \tfrac{1}{2}\mu_0 e_l\int \mathbf{r} \times \mathbf{u}\, dl$$

where e_l is the charge per unit length of the orbit. Similarly the angular momentum is the mass times the vector product of the radius vector and the velocity $\mathbf{u}$ or in terms of the mass per unit length of the orbit

$$\mathbf{p} = m_l\int \mathbf{r} \times \mathbf{u}\, dl$$

The two integrals are seen to be the same, and if e_l/m_l is the same as that for a localized electron it is equal to $-e/m$ or $\mathbf{m} = (-\mu_0 e/2m)\mathbf{p}$ and Eq. (10.14) becomes

$$\boldsymbol{\omega} = \frac{\mu_0 e}{2m}\mathbf{H} = \frac{e}{2m}\mathbf{B} \tag{10.15}$$

The precession is therefore equal to this expression for a circulating electron or for any system composed of circulating electrons. It is seen to be one-half the magnitude of the characteristic angular rate of rotation of a free electron moving in a magnetic field as given by Eq. (9.5′). From many lines of evidence in the field of atomic physics it appears that

an electron is actually possessed of an intrinsic spin and magnetic moment such that its precession in an induction **B** is twice that given by Eq. (10.15), *i.e.*, $\boldsymbol{\omega}_s = (e/m)\mathbf{B}$ where $\boldsymbol{\omega}_s$ is the rate of precession of the axis of spin of a free electron in the induction **B**. This quantity has not been directly observed for a free electron, but further evidence that it has this value is educed in Sec. 11.1. In the average atom the spin and orbital angular momenta interact with one another in such a way as to produce precessions in a magnetic field which are in general simple fractions of that predicted by the above equation. Hence for an atom the coefficient of **B** is written $g(e/2m)$, where g is a simple but generally improper fraction known as the *Landé g factor*. A further account of this factor and its significance will be found in treatises on atomic magnetism.[1]

One of the most important applications of the preceding theory is in the analysis of the *Zeeman effect*. This is an effect observed in the radiation from an atom in a magnetic field. Each line in the characteristic spectrum of an element is observed in most cases to split into a number of closely spaced lines known as a multiplet. The position of a line in a spectrum is a measure of the characteristic frequency of the radiation from the atom, and by the Bohr frequency relation the emitted frequency ν is equal to a constant times the change in energy content of the atom during the emission process. This is written

$$\nu = \frac{E_i - E_f}{h}$$

where E_i is the initial, E_f the final energy, and h is a constant known as Planck's constant (Sec. 6.5). Now the energy of the atom is affected by the presence of the magnetic field. This change in energy can be written

$$\delta E = \mathbf{m} \cdot \mathbf{H} = \frac{ge}{2m}\mathbf{p} \cdot \mathbf{B}$$

Thus, writing E' for $E + \delta E$, the frequency emitted by an atom in a magnetic field is

$$\nu = \frac{(E_i' - E_f')}{h} = \nu_0 + \frac{e}{2mh}(g_i\mathbf{p}_i - g_f\mathbf{p}_f) \cdot \mathbf{B}$$

where ν_0 is the frequency emitted in the absence of a field. A further discussion of the effect would require an introduction of the modern quantum theory. However, in simple instances the g's are equal to unity and the initial and final components of the angular momentum along the field only differ by $\pm\frac{h}{2\pi}$. In this case the equation reduces to

$$\nu = \nu_0 \pm \frac{eB}{4\pi m}$$

and a spectrum line is accompanied by two additional lines which differ from it in frequency by $\pm\frac{eB}{4\pi m}$. This is known as the simple Zeeman effect. In general, the

[1] VANVLECK, "The Theory of Electric and Magnetic Susceptibilities," Oxford University Press, New York, 1932; STONER, "Magnetism and Matter," Methuen & Co., Ltd., London, 1934.

pattern observed is quite complex and is further complicated by the fact that the g's are themselves found to be functions of the magnetic induction B. A further analysis of the effect is beyond the scope of this treatment but will be found in the references that have been cited.

Nuclear Induction.—Further very interesting and important information can be obtained about magnetic moments such as those of paramagnetic atoms and molecules (Sec. 11.5) or atomic nuclei that retain their identity in varying magnetic fields. If a fluctuating magnetic field is applied to them, these systems react like damped resonant mechanical systems and the sign and magnitude of the ratio m/p can be determined. A requisite for obtaining simple interpretable results is that this ratio, which will be written as γ for convenience, shall be invariant under the existing conditions and that the damping brought about by coupling with the rest of the atom or neighboring atoms shall not be so great as to obscure the resonance effects completely. These conditions are readily achieved for atomic nuclei that are but loosely coupled in orientation with the atom as a whole and for which γ is invariant.

The basic mechanical equation is that of the gyroscope subjected to a torque normal to its axis of rotation. Neglecting all but magnetic forces, the torque $\mathbf{T}$ is $\mathbf{m} \times \mathbf{H}$ and the rate of change of angular momentum is $d\mathbf{p}/dt$. Writing $\mathbf{p} = \mathbf{m}/\gamma$ the equation is

$$\frac{d\mathbf{m}}{dt} = \gamma \mathbf{m} \times \mathbf{H} \tag{10.16}$$

Let $\mathbf{H}$ be the sum of two fields, $\mathbf{H}_0$ which is large and constant and $2\mathbf{H}' \cos \omega t$ where $\mathbf{H}'$ is a small vector at right angles to $\mathbf{H}_0$. The latter can be written as the sum of two fields $\mathbf{H}_+ = \mathbf{H}'e^{j\omega t}$ and $\mathbf{H}_- = \mathbf{H}'e^{-j\omega t}$ by writing $\cos \omega t$ in terms of its complex components to bring out the fact that it can be considered as the sum of two equal vectors rotating in positive and negative senses about $\mathbf{H}_0$ (see also Sec. 12.7). From Eq. (10.16) it is evident that $d\mathbf{m}/dt$ is perpendicular to both $\mathbf{m}$ and $\mathbf{H}$; thus $\mathbf{m}$ does not change in magnitude but only in direction. It may be shown that a solution of Eq. (10.16) exists for which $\mathbf{m} = \mathbf{m}_0 + \mathbf{m}'e^{j\omega t}$ where $\mathbf{m}_0$ is parallel to $\mathbf{H}_0$ and $\mathbf{m}'e^{j\omega t}$ is parallel to $\mathbf{H}_+$. If the assumed values of $\mathbf{m}$ and $\mathbf{H}$ are substituted in Eq. (10.16), the result is seen on multiplying by $e^{-j\omega t}$ to be

$$j\omega \mathbf{m}' = \gamma(\mathbf{m}_0 \times \mathbf{H}' + \mathbf{m}' \times \mathbf{H}_0 + \mathbf{m}_0 \times \mathbf{H}'e^{-2j\omega t})$$

The senses of the vectors are shown in Fig. 10.16a. The left-hand side of the equation represents a vector $\pi/2$ ahead of $\mathbf{H}_+$ in rotation about $\mathbf{H}_0$, the next two vectors on the right are constant in this rotating frame, and the third vector on the right fluctuates so rapidly with time that its

average effect in the case of small $\mathbf{H}'$ is negligible. Writing $\mathbf{H}^*$ for the field for which the normal unforced precession would be ω, the equation for the magnitudes of $\mathbf{m}_0$ and $\mathbf{m}'$ is

$$m'\delta = m_0$$

where $\delta = (H_0 - H^*)/H'$. Recalling that the sum of the squares of m_0

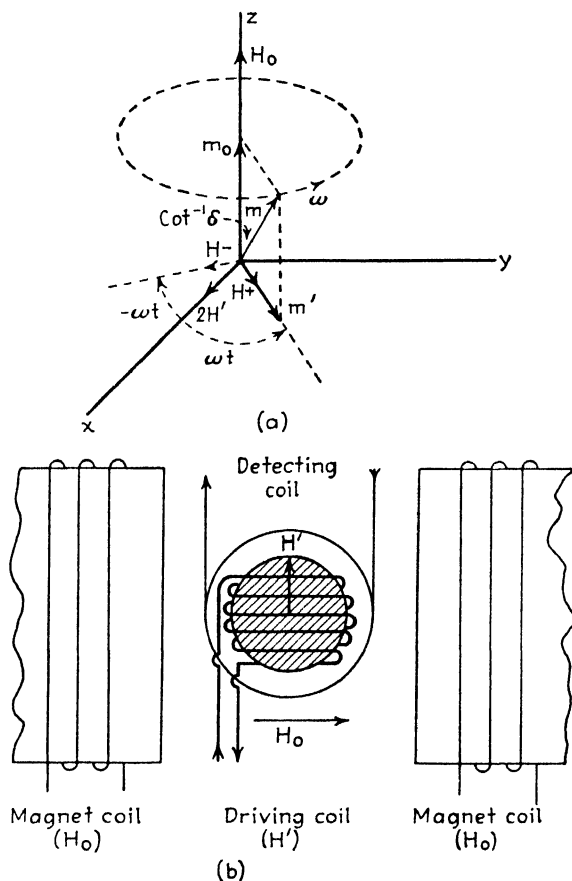



FIG. 10.16.—Nuclear induction. (a) Representation of magnetic field and magnetic moment vectors. (b) Schematic depiction of disposition of fields and coils about the sample.

and m' is the square of the constant m

$$m_0 = \frac{\delta}{\sqrt{1 + \delta^2}} m, \qquad m' = \frac{1}{\sqrt{1 + \delta^2}} m \tag{10.17}$$

These are the components of a vector $\mathbf{m}$ making an angle $\cot^{-1}\delta$ with $\mathbf{H}_0$ rotating about this axis with the angular velocity ω in such a phase as to be in the plane containing $\mathbf{H}_0$ and $\mathbf{H}_+$. If the assumption had been made

that $\mathbf{m} = \mathbf{m}_0 + \mathbf{m}'e^{-j\omega t}$, it can readily be seen that Eqs. (10.17) again result, but in this case $\mathbf{m}$ rotates in the negative sense in such a way as to lie in the plane of $\mathbf{H}_0$ and $\mathbf{H}_-$. The sense of rotation of $\mathbf{m}$ about $\mathbf{H}_0$ is determined by γ; if negative, the rotation is in the assumed sense of ω; and if positive, it rotates in the opposite sense. If ω is gradually increased, $\mathbf{H}^*$ increases and the angle between $\mathbf{H}_0$ and $\mathbf{m}$ increases until at resonance $\mathbf{m}_0$ vanishes; and as ω is further increased, $\mathbf{m}_0$ becomes negative. The same effect may be brought about by decreasing $\mathbf{H}_0$. From an experiment determining the resonant condition between ω and $\mathbf{H}_0$, *i.e.*, $\mathbf{H}_0 = \mathbf{H}^*$, and the sense of rotation of $\mathbf{m}$ the magnitude and sign of γ can be determined.[1]

Experiments of this resonance-precession type were first performed in 1945.[2] One group of experimenters observed the resonant condition by the reaction on the alternating-current circuit producing $\mathbf{H}'$ when at resonance the maximum energy is transferred from this circuit to the sample of material through the medium of the rotating nuclear moments. The other group observed the induction of an alternating field at right angles to $\mathbf{H}_0$ and $\mathbf{H}'$ brought about by the rotation of the nuclear moments. The disposition of the magnet poles producing the field $\mathbf{H}_0$ and of the high-frequency driving and detecting coils for the second type of experiment is shown schematically in Fig. 10.16*b*. In the absence of any resonant precession the magnetic field has no component normal to the detecting coil and no signal is observed in it. As resonance is approached by varying $\mathbf{H}_0$ or $\mathbf{H}'$, the precessing moments induce a characteristic signal in the detecting coil at right angles to the driving one. To determine γ from the phase of this signal care must be taken to pass through resonance from the previous equilibrium condition in the proper sense. The nature of the signal depends upon the rapidity with which the resonant condition is traversed. The exchange of energy between the nuclei and the surrounding electronic structure tends to destroy the ideal phase relationship which gives rise to a maximum signal just at resonance. For further information on the measurement of nuclear and paramagnetic atomic moments the original literature should be consulted.[3]

Atomic and nuclear magnetic moments can also be measured under conditions such that the interaction between neighboring atoms is negligible using a technique originated by Stern and Gerlach. In the hands of these investigators and Rabi and his group it has yielded valuable information in this important field. A beam of atoms or molecules traverses a long evacuated path through an inhomogeneous

[1] BLOCH, *Phys. Rev.*, **70**, 460 (1946).

[2] PURCELL, TORREY, and POUND, *Phys. Rev.*, **69**, 37 (1946); BLOCH, HANSEN, and PACKARD, *Phys. Rev.*, **69**, 127 (1946).

[3] BLOCH, HANSON, and PACKARD, *loc. cit.;* PURCELL, TORREY, and POUND, *loc. cit.;* ZAVOISKY, *J. Phys. U.S.S.R.*, **9**, 211, 245, 447 (1945); ARNOLD and ROBERTS, *Phys. Rev.*, **70**, 766 (1946); CUMMEROW, HALLIDAY, and MOORE, *Phys. Rev.*, **72**, 1233 (1947).

magnetic field. The field deflects the beam and this deflection is a measure of the magnetic moments of the particles. The apparatus is illustrated schematically in Fig. 10.17. The substance to be investigated is placed in an oven, in the side of which is a small orifice. When the oven is heated, the substance is vaporized and a beam of atomic vapor issues from the opening. The beam is defined by a series of slits and passes through a long region of inhomogeneous field before impinging on the detecting plate. The actual method of detection depends on the substance being investigated. In the case of hydrogen a plate of molybdenum oxide can be used; the hydrogen reduces the oxide and leaves a trace of metallic molybdenum where it strikes. The field may be produced by currents or permanent magnets; a section through representative pole pieces is also shown in the sketch. The field is symmetrical about a vertical plane including the beam and in the region of the beam the only appreciable components of **H** and its gradient are in the vertical direction. From the preceding chapter it will be recalled that the force on a magnet of moment **m** in a field **H** is (**m** · **grad**) **H**, or choosing x as the vertical coordinate, only H_x and its partial derivative with respect to x do not vanish and the force on an atomic magnet is $m_x\, \partial H_x/\partial x$.

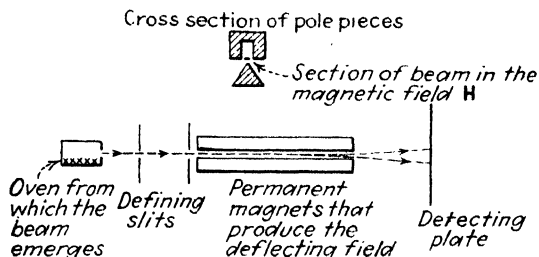



FIG. 10.17. -Schematic diagram of a Stern-Gerlach apparatus for measuring atomic magnetic moments.

Since the displacement of the beam is proportional to this force if $\partial H_x/\partial x$ is known, m_x can be found. Many atomic moments have been determined in this way and it has been shown that the component in the direction of the field cannot have any arbitrary value but is limited to an integral multiple of $\mu_0 geh/4\pi m$. The quantity $\mu_0 eh/4\pi m$ is the magnetic moment of a normal hydrogen atom and is known as the *Bohr magneton.* The accuracy of the original method of Stern and Gerlach is severely limited by the Maxwellian distribution of velocities among the atoms in the beam. As the actual displacement of an atom is inversely proportional to its velocity, a large range of velocities produces a broad trace. Many ingenious variations due principally to Rabi and his associates have greatly mitigated this disadvantage. The method has been refined to such an extent that even nuclear moments, which are about one-thousandth of the moments associated with the electron structure, can be measured to a high degree of precision. It has been mentioned previously that the value of γ for the intrinsic moment and spin of an electron is e/m. Molecular beam and nuclear induction experiments have both shown that the analogous ratio for the simplest nucleus, that of hydrogen, is not an integral multiple of $e/2M$, where M is the proton mass, but 2.7896 times this quantity. For a further account of this important technique and the results that have been attained with it the original references and review articles should be consulted.[1]

[1] ESTERMANN, *Rev. Mod. Phys.*, **18,** 300 (1946); KELLOGG and MILLMAN, *Rev. Mod., Phys.* **18,** 323 (1946).

Problems

1. Show that the electric field induced in a continuous medium by a changing induction is equal to minus the partial derivative of the vector potential with respect to the time.

2. A magnetic field **H** is parallel to the axis of a cylinder of radius a and dielectric constant κ. If the cylinder is rotated about its axis with an angular velocity ω (inducing an electric field of $\mathbf{u} \times \mathbf{B}$), show that the resultant polarization per unit volume is given by $\frac{(\kappa - 1)H\omega}{c^2}\mathbf{r}$ and that a charge equal to $\frac{2\pi a^2(\kappa - 1)H\omega}{c^2}$ appears on the surface of the cylinder per unit length. (c^2 is written for $1/\kappa_0\mu_0$.)

3. A brass disk is mounted on an axle parallel to a magnetic field H. A current i flows to the disk through a contact on the periphery and away from it along the center of the axle (Faraday disk). Show that the torque exerted on the disk is $\mu_0 \mathrm{H} i a^2/2$, where a is the radius of the disk.

4. The rails of a railway track are 1.5 m. apart and assumed to be insulated from one another. If they are connected together through a millivoltmeter, what is the reading of the instrument when a train is passing at 100 km. per hour? Assume that the vertical component of the induction due to the earth's field is 0.15 gauss.

5. Explain qualitatively why a copper sheet in falling through a horizontal magnetic field appears to be moving in a very viscous medium. How could it be distinguished from an ordinary viscous medium by changing the orientation of the sheet with respect to the field?

6. A circular coil of radius r composed of n turns of wire is suspended in a uniform magnetic field by a fiber with a restoring torque proportional to the sine of the angular displacement. If the plane of the coil with no current flowing through is parallel to the field, find its angular displacement when it carries a current i.

7. Show that if the coil of a tangent galvanometer is rotated about a vertical axis so as to lie in the same plane as the needle when the deflection is read the current is given by

$$i = \frac{2bH_e}{n} \sin \theta$$

where b is the radius of the coil, n the number of turns it contains, H_e the earth's magnetic field, and θ the angular deflection. The instrument used in this way is known as a sine galvanometer. Compare the incremental sensitivities, $d\theta/di$, for the sine and tangent galvanometers.

8. The coil of a D'Arsonval galvanometer has 100 turns and is 2 cm. on a side. The magnetic induction through the gap is 1,000 gauss (0.1 weber per square meter) and the restoring torque is 10^{-8} gm. cm. per degree. Find the angular deflection of the instrument per milliampere. If the angular deflection is measured by the deflection of a beam of light, what is the current corresponding to a deflection of 1 mm. on a scale at a distance of 2 m.?

9. Determine the shunt and series resistances that would be used to build a multi-range ammeter and voltmeter using a galvanometer movement of sensitivity 10^{-6} amp. per division and resistance of 100 ohms. The scale has 100 divisions and the ranges required are: 0–10^{-3}, 0–1, and 0–10 amp. and 0–1 and 0–100 volts.

10. The internal resistance of an ammeter is R. What resistance must be inserted in parallel to multiply the range of the instrument by n? The internal resistance of a voltmeter is R; what resistance must be included in series to multiply the range of the instrument by n?

11. A tangent galvanometer having a period of 8 sec. deflects 15° when a steady current of 0.1 amp. passes through it. Assuming negligible damping what charge will produce a throw of 30° when the instrument is used ballistically?

12. The coil of a D'Arsonval galvanometer has a period of 10 sec. for free oscillation and after a deflection successive maxima on one side are observed to be in the ratio of $\frac{10}{11}$. A steady current of 10^{-4} amp. produces a deflection of 50 scale divisions. What is the capacity of a condenser which when charged to a potential of 100 volts and discharged through the instrument produces a throw of 8 divisions?

13. A circular coil of n turns and radius b rotates with an angular velocity ω about a diameter perpendicular to a uniform magnetic field H. Show that the current flowing at any instant in the coil is

$$\frac{\mu_0 \pi n \omega b^2 H}{(R^2 + \omega^2 L^2)^{1/2}} \sin(\omega t - \phi)$$

where R and L are the resistance and self-inductance, respectively, of the coil and $\phi = \tan^{-1}(\omega L/R)$. What is the angle between the normal to the plane of the coil and the field at maximum current?

14. Neglecting friction show that the average value of the torque necessary to maintain the rotation of the preceding problem is

$$\frac{n^2 \phi_m^2 \omega R}{2R^2 + 2\omega^2 L^2}$$

where ϕ_m is the maximum flux through the coil. Find the angle between the normal to the coil and the field when the maximum torque occurs.

15. Show that in Prob. 13 the average components of the magnetic field at the center of the coil which are parallel and perpendicular to H are

$$H\left(1 - \frac{\mu_0 \pi n^2 \omega^2 L b}{4(R^2 + \omega^2 L^2)}\right) \quad \text{and} \quad \frac{\mu_0 \pi n^2 \omega b R H}{4(R^2 + \omega^2 L^2)}$$

respectively. Hence, if the velocity of rotation is small enough to neglect ωL in comparison with R, the average change in the direction of the magnetic field at the center of the coil due to its rotation is

$$\tan^{-1} \frac{\pi n^2 \omega b \mu_0}{4R}$$

This method is used in making an absolute determination of R by mounting a small test magnet at the center of the coil and noting its deflection.

16. If a small circuit of area A is placed at the center of the rotating coil of Prob. 13 with its plane normal to H, show that the emf. induced in it is

$$\frac{\mu_0^2 n^2 \pi \omega^2 A b H}{2(R^2 + \omega^2 L^2)^{1/2}} \cos(2\omega t - \phi)$$

17. Two circuits with coefficients of self-inductance L_1 and L_2 and resistances R_1 and R_2 lie near each other. If the coefficient of mutual inductance between them is L_{12}, show that a quantity of charge equal to VL_{12}/R_1R_2 will be caused to circulate through one of them if a battery of potential V is suddenly connected in series with the other.

18. A superconducting ring which is constrained to move in a vertical direction lies on a table over a coil of wire. If a current i is sent through the coil, show that the

ring will rise to a maximum height

$$h = \frac{1}{2}\frac{i^2}{mg}\frac{L_0^2 - L_h^2}{L_r}$$

where L_0 and L_h are the coefficients of mutual inductance between the coil and ring initially and at the height h, respectively; L_r is the self-inductance of the ring; m is its mass; and g is the acceleration of gravity. Describe the subsequent motion.

19. A current is induced in a coil A by a current $i_0 \sin \omega t$ in a coil B. Show that the mean force tending to increase any coordinate x specifying the position of A is

$$-\frac{1}{2}\frac{i_0^2\omega^2 L_A L_{12}}{R^2 + \omega^2 L_A^2}\frac{\partial L_{12}}{\partial x}$$

where R and L_A are, respectively, the resistance and self-inductance of the coil A and L_{12} is the mutual inductance between the coils.

20. A parallel-plate condenser consists of two circular plates of area A. If the charge on the plates fluctuates periodically and is given by $q = q_0 \sin \omega t$, show that the magnetic field between the plates due to the displacement current is given by

$$H = \frac{r\omega q_0}{2A}\cos \omega t$$

where r is the radial distance from the axis of symmetry of the plates (edge effects are neglected).

21. A current flows in a straight wire of circular cross section, and the current density i_v is parallel to the axis of the wire and a function only of the distance from the axis and the time. That is, i_v is $i_v(r, t)$ and the lines of induction are coaxial circles. Neglecting the displacement current, show that

$$\frac{\partial}{\partial r}(Hr) = ri_v \qquad \text{and} \qquad \frac{\partial E}{\partial r} = \mu_0\frac{\partial H}{\partial t}$$

where H is the azimuthal field and E and i_v are the axial electric field and current density, respectively. [Use Eqs. (9.18) and (10.3).]

22. Assuming that the conductivity of the wire in the previous problem is σ, show that the current density obeys the equation

$$\frac{\partial}{\partial r}\left(r\frac{\partial i_v}{\partial r}\right) = (j\omega\mu_0\sigma)ri_v$$

if i_v is any function of r times the periodic term $e^{j\omega t}$. Assuming that $(\mu_0\sigma\omega r^2)$ is small for any value of r in the wire, show that

$$i_v = Ae^{j\omega t}\left(1 + \frac{j\mu_0\sigma\omega r^2}{4}\right)$$

is an approximate solution. Find the real value of the current density.

23. If I is the total effective current (the root-mean-square value of the integral of i_v over the area), show that the heat developed in the wire per unit length per unit time is

$$P_l = \frac{I^2}{\pi a^2\sigma}\left[1 + \frac{1}{192}(\mu_0\sigma\omega)^2a^4\right]$$

where a is the radius of the wire. Find the fractional difference in alternating- and direct-current resistance.

24. A wire of mass ρ per unit length is stretched tautly between two fixed pegs a distance l apart. If the wire is plucked, it vibrates at its fundamental frequency ω_0 and the amplitude of vibration falls to $1/e$ of its initial value at the end of τ sec. The ends of the wire are then connected through a circuit of total resistance R, and the wire is placed in a uniform magnetic field $H = B/\mu_0$ normal to its length. Assuming that the wire vibrates in the form of a half sine curve between the pegs at adjacent nodes, show that the time τ' for it to fall to $1/e$ of its maximum amplitude after being plucked is given by

$$\frac{1}{\tau'} = \frac{1}{\tau} + \frac{B^2 l}{\pi R \rho}.$$

If an alternating current i is sent through the wire, find the frequency that will make the amplitude of vibration a maximum and the value of this maximum in terms of i.

25. A beam of molecules of magnetic moment m enters a region of constant magnetic field. The beam axis is normal to the field and makes an angle θ with the normal to the plane bounding the region of uniform magnetic field at entrance. Assuming that the molecules take up positions parallel or antiparallel to the field after crossing the boundary, show that the beam will be deflected at the boundary through a small angle δ normal to the field where $\delta = \pm(mH/2U) \tan \theta$. The quantity U is the kinetic energy of the molecules in the entering beam, and it is assumed that $U >> mH$.

CHAPTER XI

MAGNETIC PROPERTIES OF MATTER

11.1. Magnetomechanical Effects.—There are a number of phenomena which indicate that the magnetic properties of matter in bulk are due to the circulatory or spinning motion of the electrons contained in it. Though the atomic nucleus is responsible for most of the mass, its contribution to the magnetic characteristics of an atom is very small. However, the forces acting between the electrons and the nucleus serve to transfer electromagnetic reactions to matter in bulk. The three-dimensional lattice characteristic of a crystal is defined by the regular spatial arrangement of these nuclei and the electrons which are more closely bound to them. In metals certain of the electrons are comparatively free to move through the lattice under the influence of electric fields, but even those conduction electrons are not entirely free from the influence of the lattice, as is shown by the phenomenon of resistance. There is a continuous dissipation of energy at the rate $\mathbf{E} \cdot \mathbf{i}_v$ per unit volume in a conductor which represents a transfer of energy from the conduction electrons to the comparatively rigid atomic systems in their path. The same thing is brought out by the force on a current-carrying conductor in a magnetic field. The force is brought into existence by the relative motion of the conduction electrons and the field, but because of the interaction between these and the ion-lattice structure the force is transmitted to the crystal as a whole. Though there is some tendency for the conduction electrons to move to one side in a field (Hall effect), it is relatively small and the current is almost uniformly distributed over the cross section of the conductor.

An interesting experiment due to Tolman and Stewart, which depends on this interaction, shows not only that the carriers of electricity in a metal have a negative ratio of charge to mass but also gives a rough measure of this ratio. A coil of wire is wound on the periphery of a disk which is mounted on an axle. The ends of the coil are brought out through slip rings on the axle to the terminals of a galvanometer. The coil is set in rapid rotation about the axle and then brought to rest as suddenly as possible. The interval of deceleration should be small in comparison with the period of the galvanometer so that the instrument will behave ballistically. The momentum of the conduction electrons carries them on after the coil has stopped and this surge of current produces a ballistic throw. The force acting on an electron which is equal to its rate of change of momentum can be written $E'e$, where E' is an electric field which would produce the same motion and e is the electronic charge. But E' is the effective potential difference between the ends divided by the length of the wire, or

from Eq. (10.2)

$$\frac{d(mv)}{dt} = E'e = \frac{e}{l}\left(L\frac{di}{dt} + Ri\right)$$

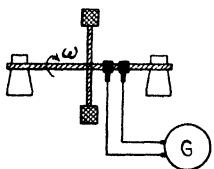
where l is the length of the wire and L and R the self-inductance and resistance, respectively, of the circuit. This equation can be multiplied by dt and integrated over the small time interval δt during which the deceleration takes place. The left-hand side is then equal to the electron's momentum at the beginning of the interval since it vanishes at the end; the first term on the right vanishes as i is zero at both the beginning and end of the deceleration and the last term contributes eRq/l, where q is the circulated charge. Writing the result explicitly in terms of e/m

$$\frac{e}{m} = \frac{lv}{Rq}$$

FIG. 11.1.—Schematic diagram of the Tolman and Stewart apparatus for determining e/m for the carriers of electricity in a metal.

The initial velocity v is equal to $b\omega$, where b is the radius of the coil and ω its initial angular velocity and q is given in terms of the throw by Eq. (10.11); therefore neglecting damping

$$\frac{e}{m} = \frac{2\pi lb\omega}{RS_{T_f}\theta_m}$$

All the quantities on the right are known, and though the experiment is not capable of great accuracy, the results obtained show that e/m is negative and equal approximately to 4.5×10^{17} esu. per gram or 1.5×10^{11} coulombs per kilogram. This is equal (within the limits of accuracy of the experiment) to the value found for free electrons.

Since magnetic effects are presumably due to the spin and circulation of all the electrons in a sample of matter, not merely those associated with conduction, it is reasonable to suppose that there should be a pronounced torsional reaction upon the orientation of the electron axes necessary to produce a magnetic moment. This type of magnetomechanical effect has been observed and its investigation has yielded important information regarding the nature of magnetism. Two converse effects should evidently be expected: (1) if a specimen of matter is rotated, a magnetic moment should be induced; and (2) if a magnetic field is established near the specimen, the latter should experience a mechanical torque. The first of these is known as the *Barnett* effect and the second as the *Einstein-de Haas* effect. Both have been observed.[1] Consider first the Barnett effect. The rotation of a cylinder with an angular velocity ω in a field-free region should have the same effect on the electron circulation and spin responsible for magnetism as the application of a magnetic field equal to $-\omega p/m$ [Eq. (10.14)]. The magnetic moment induced in the specimen by a field **H** can be measured and like-

[1] For a detailed account of the gyromagnetic and electron inertia effects see Barnett, *Rev. Mod. Phys.*, **7**, 129 (1935).

wise the moment induced by the rotation ω. If $\mathbf{H}$ is the magnetic field that induces the same moment as the rotation ω, $\mathbf{H} = -\omega p/m$. From a knowledge of $\mathbf{H}$ and ω the ratio of the magnetic to mechanical moment of the entities responsible for magnetism can be determined. If this is written in the notation of Sec. 10.6,

$$\frac{\omega}{\mathbf{H}} = \frac{\mu_0 ge}{2m}$$

The effect is large enough to be measured only in the case of ferromagnetic materials. By using this procedure Barnett found that for iron $g = 1.929$ and for the alloy permalloy $g = 1.906$. Thus the value of g associated with these materials is very close to 2 which is the ratio characteristic of a spinning electron. This is evidence that magnetism, in ferromagnetic substances at least, is largely due to the effects of electron spin. However, the values of g obtained by Barnett are accurate to within about 0.5 per cent and it is seen that the observed values of g differ from 2 by amounts greater than the probable error of the experiment. So the effects of orbital electron rotation are appreciable in these substances.

The Einstein-de Haas effect can be used to measure the effective value of g for elements other than iron and its alloys. One method is to suspend a sample of the material in the form of a cylinder about 0.03 cm. in diameter and 10 cm. long by means of a fine quartz fiber along the axis of a vertical solenoid. The sample carries a mirror for observing its deflection. If a current i is sent through the soleniod, a field $n'i$, where n' is the number of turns per unit length, is established. From the conservation of angular momentum and Eq. (10.14) this is equivalent to the transmission of an initial angular momentum to the sample. As in the case of the ballistic galvanometer, the throw of the sample is proportional to this initial angular momentum. If damping is negligible, energy is conserved, and equating the initial kinetic energy $\frac{1}{2}I\omega_0^2$ and the potential energy stored in the fiber at the maximum deflection, $\frac{1}{2}k\theta_m^2$

$$\theta_m = \left(\frac{I}{k}\right)^{\frac{1}{2}} \omega_0 \qquad \text{or} \qquad P = I\omega_0 = \frac{\tau}{2\pi} k\theta_m$$

where τ is the period of oscillation of the sample. The magnetic moment M can be measured by methods that will be described later in this chapter. From a knowledge of M and a measurement of k, τ, and θ_m, the ratio of M to P can be determined. The method is suitable for materials in which large magnetic moments are induced. For iron and nickel M/P is found to be about 1.8 $(\mu_0 e/2m)$ in agreement with the results of Barnett. For most substances the magnetic moment induced is too small for the measurement of either P or M to be made in this way.

However, the effect can be magnified by means of a resonance technique, and if the ratio between the induced magnetic moment per unit volume and the inducing field is known, the effective value of g can be determined. The equation of rotation of the sample is analogous to that of a galvanometer coil. The torque is dP/dt or if the ratio of the magnetic to mechanical moment is $\mu_0 ge/2m$, $T = \frac{2m}{\mu_0 ge}\frac{dM}{dt}$. From the galvanometer discussion the equation of motion can be written

$$\frac{d^2\theta}{dt^2} + \frac{2\delta}{\tau}\frac{d\theta}{dt} + \omega_0^2\theta = \frac{1}{I}\frac{2m}{\mu_0 ge}\frac{dM}{dt}$$

where δ is the logarithmic decrement, τ is the free period, ω_0 is the associated natural angular velocity, and I is the moment of inertia of the sample. Let an alternating current be applied to the solenoid so that the field surrounding the sample is given by the real part of $H_0e^{j\omega t}$. If the induced moment per unit volume per unit field of the material of the sample is χ_m (volume susceptibility) then $M = \chi_m VH_0e^{j\omega t}$. The steady motion of the system will be periodic with the impressed period, *i.e.*, $\theta = \theta_m e^{j\omega t}$, and substituting these expressions in the above equation

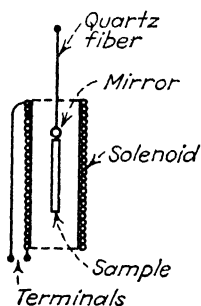



FIG. 11.2.—Section through an apparatus for detecting the torsional impulse associated with magnetization.

$$\left(-\omega^2 + \frac{j\omega 2\delta}{\tau} + \omega_0^2\right)\theta_m = \frac{2m\chi_m VH_0 j\omega}{\mu_0 geI}$$

If the impressed period is equal to the natural period of the system, the first and third terms on the left cancel one another; the maximum deflection is then very large and given by

$$\theta_m = \frac{2m}{\mu_0 ge}\frac{\chi_m VH_0\tau}{2\delta I}$$

It is assumed that χ_m is known and as all the other quantities are readily measured the amplitude of oscillation of the system when the current through the solenoid has the same period as that of the suspended sample determines g for the substance. Since m/e is small, the rest of the expression should be made as large as possible. Sucksmith employed this method for certain elements in the iron and rare-earth groups. The sample was packed in a tube 0.06 cm. in diameter and 6 cm. long. The natural period of the system was 2 sec. and the damping was reduced by evacuating the surrounding region. Since disturbing effects are proportional to the square of the magnetic field, nothing is gained by increasing this factor beyond a certain point, and a value of about 5×10^4 amp.-turns per meter was used. Maximum angular deflections

of the order of 10^{-2} were obtained, which permitted an accuracy of about 5 per cent in the determination of g. Table I gives the values of g obtained by Sucksmith for various ions. The superscript indicates the valence and the quantity in brackets the number of electrons associated with the ions. It is seen that certain of the rare-earth ions have values of g even greater than for a spinning electron.[1]

TABLE I

Ion	Form of sample	Value of g
Cr^3 (21)	CrCl	1.95
Mn^2 (23)	$MnCO_3$ & $MnSO_4$	1.98
Fe^2 (24)	$FeSO_4$	1.89
Co^2 (25)	$CoSO_4$ & $CoCl_2$	1.54
Nd^3 (57)	Nd_2O_3	0.78
Eu^3 (60)	Eu_2O_3	>4.5
Gd^3 (61)	Gd_2O_3	2.12
Dy^3 (63)	Dy_2O_3	1.36

11.2. General Theory of Magnetic Materials.—The preceding discussion of the mechanical reaction of matter to the establishment of a magnetic field in its neighborhood strongly supports the view that the magnetization of matter is due to the induced precession of the circulating or spinning electrons that it contains. The majority of materials are isotropic and the axis of precession is determined by the externally applied magnetic field, the substance acquiring a net magnetic moment parallel to this direction. These circulating currents are known generally as *amperian currents* and they resemble the persistent currents in superconductors rather than ordinary currents in that their flow involves no energy dissipation. The electrical evidence for the existence of these

[1] Recent experiments of the type described in Sec. 10.7 in which ferromagnetic materials have been used confirm both the general theory given in that section and the conclusion that the entity responsible for ferromagnetic properties exhibits a value of g of the order of 2. In these ferromagnetic experiments, first performed by Griffiths [*Nature*, **158,** 670 (1946)] and reviewed by Kittel [*Phys. Rev.*, **73,** 155 (1948)], resonance absorption is evidenced by the loss of energy occurring most strongly at a critical angular frequency. A ferromagnetic sample in a magnetic field is made part of a high-frequency resonant circuit, and strong absorption of electromagnetic energy is observed in accordance with the general theory of Sec. 10.7 and at the frequency predicted by Eq. (10.16). For reasons indicated in subsequent sections of this chapter the geometry of the sample influences the resonant frequency through the dependence of effective field on sample shape. But with due regard to this factor it is found that the apparent value of g is close to 2. The actual values observed are somewhat in excess of 2 for reasons doubtless connected with the interaction of the spinning electrons with the crystal lattice, although this is not as yet understood quantitatively.

currents comes only from the magnetic moment to which they give rise, and in dealing quantitatively with the magnetic properties of matter it is more convenient to express the energy represented by such a system in terms of its magnetic moment than in terms of the amperian currents. The general relations between the magnetic moment per unit volume of the substance and the amperian currents which give rise to it can be obtained by comparing the expressions for the energy in terms of these two quantities. The energy of a current vortex of area $\mathbf{a}$ which is placed in a region of induction $\mathbf{B}$ is by Eq. (9.20) $i\mathbf{a} \cdot \mathbf{B}$ or, since the magnetic moment associated with such a vortex is $\mathbf{m} = \mu_0 i\mathbf{a}$, the energy can be written $\mathbf{m} \cdot \mathbf{B}/\mu_0$. If $\mathbf{m}_v$ is the moment per unit volume of a substance, the energy per unit volume is $\mathbf{m}_v \cdot \mathbf{B}/\mu_0$. The total energy is obtained by integrating over the volume occupied by the matter, so in terms of the vector potential $\mathbf{A}$, defined by $\mathbf{B} = \text{curl } \mathbf{A}$, the energy can be written

$$U = \frac{1}{\mu_0}\int \mathbf{m}_v \cdot \text{curl } \mathbf{A} \, dv$$

The integrand can be rewritten by means of the vector identity

$$\mathbf{m}_v \cdot \text{curl } \mathbf{A} = \mathbf{A} \cdot \text{curl } \mathbf{m}_v + \text{div } (\mathbf{A} \times \mathbf{m}_v) \quad \text{(appendix } D\text{)}$$

and as the volume integral of the divergence is equal to the surface integral of the normal component of its argument

$$U = \frac{1}{\mu_0}\int \mathbf{A} \cdot \text{curl } \mathbf{m}_v \, dv + \frac{1}{\alpha_0}\int (\mathbf{A} \times \mathbf{m}_v) \cdot \mathbf{ds}$$

The energy can also be written, in terms of a large number of current filaments that are assumed to traverse the matter, in the form of Eq. (9.21). Since there are two terms in the previous expression for the energy, the amperian currents will also be divided into two groups, volume currents of density $\mathbf{i}_v^a$ and surface currents which flow over any bounding surfaces with a density $\mathbf{i}_s^a$. Since $i \, \mathbf{dl}$ can be written as either $\mathbf{i}_v^a \, dv$ or $\mathbf{i}_s^a \, ds$ the two terms of Eq. (9.21) are

$$U = \int \mathbf{A} \cdot \mathbf{i}_v^a \, dv + \int \mathbf{A} \cdot \mathbf{i}_s^a \, ds$$

Since these are two expressions for the same energy, the integrands can be equated yielding

$$\mu_0 \mathbf{i}_v^a = \text{curl } \mathbf{m}_v$$

and (11.1)

$$\mu_0 \mathbf{i}_s^a = \mathbf{m}_v \times \mathbf{n}$$

as $\mathbf{ds} = \mathbf{n} \, ds$, where $\mathbf{n}$ is a unit vector perpendicular to the area ds. These equations are the relations between the amperian currents and

the magnetic moment per unit volume of the substance. In a sense they are the analogues of Eqs. (2.19) and (2.20) which give the relations between the induced charge densities and the polarization or electric moment per unit volume.

These equations also play a similar role in the theory of magnetism that Eqs. (2.19) and (2.20) play in the theory of electrostatics. Those equations enabled us to write the fundamental differential equation of electrostatics in the presence of dielectrics, div $\mathbf{D} = q_v$, in terms of the free charges q_v. In the same way these equations enable us to write the magnetic equations in the presence of magnetic materials in terms of the applied currents. In Sec. 9.3 the magnetic field was defined by Eq. (9.7) in which i represented the currents induced to flow in conductors under the influence of emfs. The fundamental vector **B** is defined as the force per unit current element, and i in Eq. (9.3) must be interpreted as all currents of whatever nature. Thus writing i_v^t for the total volume current density, Eqs. (9.3) and (9.8) become

$$\mathbf{B} = \frac{\mu_0}{4\pi}\int \frac{\mathbf{i}_v^t \times \mathbf{r}_1}{r^2}\,dv \qquad \text{and} \qquad \mathbf{H} = \frac{1}{4\pi}\int \frac{\mathbf{i}_v \times \mathbf{r}_1}{r^2}\,dv$$

By the arguments of Sec. (9.4) these expressions are equivalent to Eq. (9.18) and its analogue

$$\text{curl } \mathbf{B} = \mu_0 \mathbf{i}_v^t \qquad \text{and} \qquad \text{curl } \mathbf{H} = \mathbf{i}_v$$

Considering the total current density $\mathbf{i}_v^t$ to be made up of the currents $\mathbf{i}_v$ due to emfs. and the amperian currents $\mathbf{i}_v^a$, the equation for **B** with the help of Eq. (11.1) becomes

$$\text{curl } \frac{1}{\mu_0}(\mathbf{B} - \mathbf{m}_v) = \mathbf{i}_v$$

It is thus reasonable to identify the auxiliary vector **H** with the argument of the curl above or

$$\mathbf{H} = \frac{1}{\mu_0}(\mathbf{B} - \mathbf{m}_v) \tag{11.2}$$

The vector **B** which determines the forces experienced by a current or moving charge is not determined in general by **H** or the applied currents alone. Within magnetized matter the vector $\mathbf{m}_v$ must be known as well.

If $\mathbf{m}_v$ is given at every point, the equivalent amperian currents can be obtained from Eq. (11.1) and any problem can be solved by the methods that have been outlined in the preceding chapters. However, this method is seldom of practical importance. A more generally useful method is to take advantage of the formal analogy between magnetization and polarization and set up a scalar potential function. It was pointed

out in Sec. 9.4 that the lines of induction from an elementary current vortex or magnetic dipole are of the same form at a great distance as the lines of electric force due to an electric dipole. From Eq. (9.16) **H** can be written

$$\mathbf{H} = -\mathbf{grad}\ \Omega \tag{11.3}$$

where Ω is the magnetic scalar potential. For an elementary vortex $\Omega = \mathbf{m} \cdot \mathbf{r}_1/4\pi r^2\mu_0$ or, if a volume distribution of current vortices is to be considered

$$\Omega = \frac{1}{4\pi\mu_0}\int \frac{\mathbf{m}_v \cdot \mathbf{r}_1}{r^2} dv \tag{11.4}$$

Thus in regions outside magnetized matter the magnetic field can be calculated by the same technique as the electric field is calculated from a knowledge of the volume polarization of matter. Of course, to the value of **H** obtained in this way must be added the contribution due to applied currents. The induction **B** at any point outside matter is, of course, simply $\mu_0\mathbf{H}$, but at any point inside matter where magnetization exists $\mathbf{m}_v$ must be added to the $\mu_0\mathbf{H}$ calculated in this way in order to obtain the value of **B** as shown by Eq. (11.2). If applied currents are assumed to be absent, Eq. (11.4) shows that the problem of determining the value of **H** produced by neighboring magnetized matter is formally identical to the problem of determining **E** in the neighborhood of polarized matter. The potential Ω plays an analogous role to that played by the potential V and the magnetization $\mathbf{m}_v$ takes the place of the polarization $\mathbf{p}_v$.

The similarity between the two analyses is so striking that it suggests carrying the line of thought one step farther and introducing a magnetic quantity to describe magnetic effects in the absence of applied currents (magnetostatic effects) in the same way as electric charges are used to account for electrostatic effects. It may be pointed out that this is an unnecessary step in as much as all known magnetic phenomena are satisfactorily accounted for on the basis of Ampère's law. However, as was mentioned in Sec. 9.1, the magnetostatic effects associated with naturally occurring magnetic substances were known long before the work of Ampère, and the electrostatic method of approach is simpler though possessed of less physical reality than that of Ampère. The electric moment of two equal and opposite charges is equal to the product of the magnitude of either charge and their vector separation ($-$ to $+$). By analogy a quantity known as the *magnetic pole* can be introduced and defined by saying that the magnetic moment of two equal and opposite poles is equal to the product of the magnitude of either pole and their vector separation. Poles could equally well be designated by the symbols $+$ and $-$ and in calculation this is done; but for historical

reasons the positive pole is known as "north" and the negative one as "south." As shown in Fig. 11.3, the field in the immediate neighborhood of such a magnetic dipole is not similar to that produced by a vortex of current. This is of no importance in magnetostatic calculations since one is not interested in the actual fields inside magnetic materials but only the effective fields in which other known magnetic moments are placed. The energy associated with a magnetic dipole in a field **H** is $\mathbf{m} \cdot \mathbf{H}$ rather than $-\mathbf{p} \cdot \mathbf{E}$ as for an electric dipole for the reasons given in Sec. 9.5.

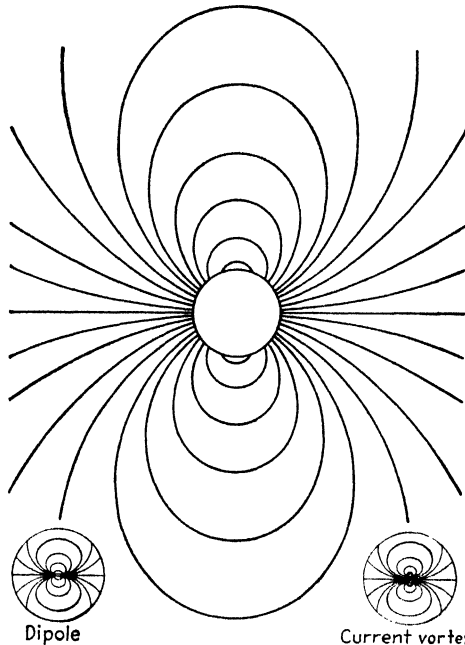



FIG. 11.3.—Lines of magnetic induction at a great distance from a magnetic dipole or current vortex.

The condition of magnetized matter can be described by giving the distribution of poles throughout its volume and over its surface in exactly the same way as the electrical condition of a polarized dielectric is described in terms of the volume and surface charge densities. By retracing the electrostatic arguments Eq. (11.4) obviously implies that the law of force between poles is the same as Coulomb's law of force between charges. If p is written for pole strength

$$\mathbf{F} = \frac{1}{4\pi\mu_0} \frac{p_1 p_2}{r^2} \mathbf{r}_1 \tag{11.5}$$

Since one north and one south pole are associated with a current, vortex poles always occur in pairs and this law cannot be verified directly but

only through the interaction of pairs of poles. Writing $\mathbf{m}$ for the magnetic moment $p\mathbf{l}$ of a pair of poles, where $\mathbf{l}$ is the vector separation, the field due to a magnetic moment $\mathbf{m}_1$ is

$$\mathbf{H} = -\mathrm{grad}\,\frac{\mathbf{m}_1 \cdot \mathbf{r}}{4\pi\mu_0 r^3} \tag{11.6}$$

and the energy of a second magnetic moment $\mathbf{m}_2$ in this field is

$$U = \mathbf{m}_2 \cdot \mathbf{H} \tag{11.7}$$

From these equations the forces and torques can be derived and the experimental verification of them is the justification for writing Eq. (11.5). Just as Eq. (1.2) defines the electrostatic system of units, Coulomb's magnetostatic law of force can be used to define what is known as the *electromagnetic system.* For this purpose the coefficient on the right is taken as unity and the unit electromagnetic pole is defined by

$$\mathbf{F} = \frac{p_1 p_2}{r^3}\mathbf{r}$$

where $\mathbf{F}$ is the force in dynes between poles of strengths p_1 and p_2 a distance r cm. apart. From this definition our line of argument can be retraced and the other magnetic and electric quantities defined. This system of units is found in many texts and for convenience a conversion table is given in the Appendix. The concept of the magnetic pole will be employed from time to time throughout this chapter and will be found particularly useful in the discussion of magnetostatics and permanent magnets. In dealing with extended bodies rather than elementary moments the word pole is used to specify a region of the surface of the body which exhibits a considerable net surface density of poles of one sign or the other. These poles are analogous to patches of charge on the surface of a dielectric. The patches of pole strength occur essentially in pairs, and in the familiar form of the horseshoe or bar magnet they are more or less localized at the extremities.

11.3. Simple Magnetic Materials.—It was seen in the previous section that a knowledge of the magnetization $\mathbf{m}_v$ is essential for analyzing the magnetic induction and determining the forces upon moving charges. If the atoms of the substance are sufficiently far apart so that their interaction is negligible or if this condition is not fulfilled but the effect of one atom upon another is of the same nature as that produced by an externally applied magnetic field, the problem is comparatively simple. It was shown in Sec. (10.7) that if a magnetic field is applied to an atomic system, the latter is induced to precess about the direction of the field. If the induced magnetic moment is proportional to the precessional angular velocity, it is both in the direction of the field

and proportional to it. Even if the atom, owing to its particular electronic configuration, has a permanent magnetic moment its average effective contribution to the magnetization will be in the direction of the field and proportional to it if the field is not too great. Thus the vectors $\mathbf{m}_v$, $\mathbf{B}$, and $\mathbf{H}$ are all proportional to one another. A simple medium of this type which is uniform throughout and without inherent directional properties is known as linear, homogeneous, and isotropic. Most materials satisfy these conditions approximately, though elements, such as iron, cobalt, and nickel, which are the most important for practical magnetic purposes, do not. These will be considered separately in a later section.

For a simple substance we can write

$$\mathbf{B} = \mu\mu_0\mathbf{H} \tag{11.8}$$

where μ is a simple constant of proportionality characteristic of the substance which is known as the *permeability*. It is the analogue of the dielectric constant κ which is characteristic of the electrical properties of a material. From Eq. (11.2) we can also write

$$\mathbf{m}_v = \frac{(\mu - 1)}{\mu}\mathbf{B} = (\mu - 1)\mu_0\mathbf{H}$$

where the coefficient of $\mu_0\mathbf{H}$ is known as the *magnetic susceptibility* of the substance and is generally written χ_m.

$$\chi_m = \mu - 1$$

If the permeability is constant throughout the entire region, any problem can be solved by the methods of Chap. IX. It is merely necessary to replace μ_0 by $\mu\mu_0$ and all forces and torques are determined. Most actual instances involve boundaries of discontinuity between media of different permeability. Here also the problem can be set up very simply in exact analogy with the similar problem in electrostatics. The argument of Sec. 9.4 shows that the vector potential $\mathbf{A}$ is a solution of the differential equation

$$\nabla^2\mathbf{A} = -\mu\mu_0\mathbf{i}_v$$

and $\mathbf{B} = \textbf{curl } \mathbf{A}$. The potential $\mathbf{A}$ that is appropriate to the problem is the solution of this equation that satisfies the necessary conditions at any boundaries. These conditions are merely that the fundamental equations

$$\text{div } \mathbf{B} = 0 \qquad \text{and} \qquad \textbf{curl } \mathbf{H} = \mathbf{i}_v$$

be satisfied. Figure 11.4 illustrates the implications of these equations at a boundary. Consider the flux of $\mathbf{B}$ through a Gaussian surface in the shape of a shallow pillbox enclosing a region of the boundary.

Since the sides perpendicular to the boundary can be made of negligible area, the flux is $\mathbf{B}_1 \cdot \delta\mathbf{s} - \mathbf{B}_2 \cdot \delta\mathbf{s}$, where the subscripts refer to the two media, and the first equation demands that this be zero. If B_n is written for the normal component of induction at the surface, the condition becomes

$$B_{n1} = B_{n2} \tag{11.9}$$

The second equation implies that the integral of $\mathbf{H} \cdot d\mathbf{l}$ around any closed path is equal to the normal current through the bounded area. The sides of the path parallel to the surface in the two media can be made very long in comparison with the portions perpendicular to the surface and the integral of $\mathbf{H} \cdot d\mathbf{l}$ becomes $\mathbf{H}_1 \cdot \delta\mathbf{l}_1 + \mathbf{H}_2 \cdot \delta\mathbf{l}_2$ or since $\delta\mathbf{l}_1 = -\delta\mathbf{l}_2$, $\mathcal{H}$ for the circuit is $(\mathbf{H}_1 - \mathbf{H}_2) \cdot \delta\mathbf{l}_1$. This is equal to the component of $\mathbf{i}_s$ normal to $\delta\mathbf{l}_1$ or $\mathbf{n} \times \mathbf{i}_s \cdot \delta\mathbf{l}_1$, where $\mathbf{n}$ is a unit vector normal to the surface. Hence

FIG. 11.4.—Boundary conditions between magnetic media.

$$\mathbf{H}_1 - \mathbf{H}_2 = \mathbf{n} \times \mathbf{i}_s$$

or multiplying each side vectorially by $\mathbf{n}$

$$(\mathbf{H}_1 - \mathbf{H}_2) \times \mathbf{n} = \mathbf{i}_s \tag{11.10}$$

which states that the difference between the tangential components of $\mathbf{H}$ on the two sides of the boundary is equal to the surface-current density. If there are no applied currents along the boundary, $H_{t1} = H_{t2}$, or

$$\frac{B_{t1}}{\mu_1} = \frac{B_{t2}}{\mu_2}$$

where μ_1 and μ_2 are the permeabilities of the two media. These conditions are seen to be closely analogous to those applying to the normal and tangential components of the electric field at the boundary between two dielectrics. If there are no applied currents in the region, $\mathbf{A}$ satisfies Laplace's equation just as V does in the absence of free charges. Thus all of the electrostatic problems of this type have exact analogues for simple magnetic substances. For example, the method of images is available for the solution of problems involving a magnetic dipole or a long straight wire carrying a current in front of the plane surface of a slab of magnetic material, just as for the analogous problems in electrostatics (however, it will be found on analysis that the image of a current does not change sign). Of course, Eq. (9.27) gives the energy associated with any configuration of currents and linear magnetic matter with the understanding that $\mathbf{B}$ and $\mathbf{H}$ are related through Eq. (11.8). The energy method is frequently the simplest one for obtaining forces and torques.

The flux density **B** in a region containing matter is not the effective flux density in which the atomic electrons circulate, for they themselves contribute to the induction at the position they occupy. The problem of finding the value of **B** effective upon an atom is closely analogous to that of finding the effective molecular electric field in a dielectric as described in Sec. 2.6. The magnetic correction is smaller than the electric one for those cases where the assumption of a linear homogeneous isotropic medium is justified. (In the case of a ferromagnetic substance where μ is large the assumptions are no longer valid.) The effective flux density **B**′ could be calculated as in Sec. 2.6 and the result written down immediately. Alternatively it may be obtained by considering that the magnetization of a block of uniformly magnetized material may be thought of as due to an infinite solenoid carrying a current i_s per unit length. The removal of a small sphere of matter (without otherwise altering the flux density) is equivalent to the creation of a hypothetical spherical cavity on the inner surface of which a current i_s flows per unit axial length in the contrary sense to that in the solenoid. This can be seen by considering sections through the

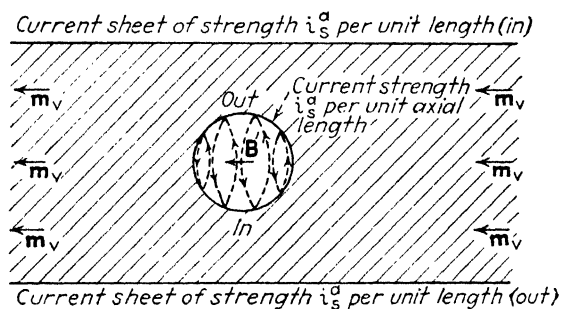



FIG. 11.5.—Local effective induction in a magnetic material.

solenoid and applying the circulation concept of Fig. 9.15. The problem of finding the induction in the cavity is similar to that of finding the induction within a uniformly charged rotating sphere (Prob. 16, Chap. IX). This additional induction is $2\mu_0 i_s/3$ or $2m_v/3$ in the opposite sense to the induction $B = \mu_v i_s$. Hence the field in the hypothetical cavity to which the atom is subject is

$$\mathbf{B}' = \mathbf{B} - \frac{2m_v}{3} = \frac{\mu + 2}{\mu - 1}\frac{\mathbf{m}}{3}$$

The actual atomic susceptibility defined as the atomic magnetic moment per unit magnetizing field, which is written α_m, is the reciprocal of the coefficient of m_v multiplied by μ_0 and divided by the number of atoms per unit volume n

$$\alpha_m = \frac{3\mu_0}{n}\frac{\mu - 1}{\mu + 2} \tag{11.11}$$

This is the quantity that is of major interest from an atomic point of view. If the influence of neighboring atoms is small, which means that μ is approximately unity, $n\alpha_m$ is seen to reduce to $\mu_0\chi_m$. From this discussion it is evident that the nature of the cavity in a block of magnetized material is of great importance in determining the effective field inside it. In the limit of a long narrow cylindrical cavity coaxial with $\mathbf{m}_v$ the field inside near the center is immediately determined by the boundary condition [Eq. (11.10)] to be **H** itself. Similarly for a shallow cylindrical pillbox coaxial with $\mathbf{m}_v$ the field inside near the center is given by the other boundary con-

dition [Eq. (11.9)] as $\mu\mathbf{H}$. The field inside cavities of other shapes is given by the solution to Laplace's equation subject to the boundary conditions imposed by the shape of the cavity. The converse problem is that of an arbitrary shaped block of magnetic material placed in an initially uniform field. These problems can not in general be solved rigorously.

11.4. Determination of μ and χ_m.—The simplest and most direct method of measuring the permeability of a substance is to choose a sample in the shape of a torus and wind a toroidal coil upon it. The field is then strictly calculable throughout the substance, and the total flux ϕ can be measured. On replacing μ_0 by $\mu\mu_0$ in the expression for the self-inductance of a toroid (Sec. 9.6), it is seen that the self-inductance, when the core has a permeability μ, is simply $\mu L'$, where L' is the self-inductance of a similar air-core toroid. Thus a measurement of the self-inductance determines μ. This is closely analogous to the determination of the dielectric constant of a material by placing it in the region between two concentric spherical conducting shells and measuring their resulting capacity. For these geometrical configurations there are no corrections to be applied. It is frequently more convenient to place the sample in a long straight solenoid, but here there are, of course, end corrections as in the case of a parallel-plate condenser. Magnetic measurements are more difficult than the analogous electric measurements, for in the case of iron and its alloys the assumption of linearity yields only a poor approximation and for other substances χ_m is so small that it is difficult to measure.

FIG. 11.6.—Apparatus for measuring the magnetic susceptibility of liquids.

One method of measuring χ_m for liquids is illustrated in Fig. 11.6. The liquid is placed in a U tube, and a uniform magnetic field is established in the region of the meniscus in one arm. The presence of the field is found to alter the level of the meniscus, and from this alteration in height the susceptibility can be calculated. The susceptibility will be assumed to be small and any variation of the field throughout the region will be neglected. Starting from Eq. (9.27) for the magnetic energy and proceeding directly in analogy with the argument in Sec. 2.5, it can be shown that the increase in energy resulting from the introduction of material of magnetic moment $\mathbf{m}_v$ into a region in which the field was previously $\mathbf{H}_0$ is given by $\frac{1}{2}\int \mathbf{m}_v \cdot \mathbf{H}_0\, dv$. Therefore in analogy with Eq. (2.37) the difference in susceptibility between the liquid χ_m and the air above it χ'_m is given by

$$\chi_m - \chi'_m = \frac{4\rho g h}{\mu_0 H_0^2}$$

Thus, if the susceptibility of air is known, the susceptibility of the liquid can be determined from a measurement of ρ, h, and the surrounding field. A somewhat analogous method for solids is that due to Curie. A small sample of the material is placed in a region of inhomogeneous field where both the field and its gradient are known. The excess energy in the volume occupied by the substance if $\frac{1}{2}(\mu - \mu')\mu_0 H^2 V$. Thus the force, say, in the x direction on the sample, is given by

$$\mathbf{F}_x = \frac{\partial U}{\partial x} = \mu_0(\chi_m - \chi_m')VH_x\frac{\partial H_x}{\partial x}$$

This force can be measured by means of a balance and the susceptibility of the substance determined if the susceptibility of air is known. These

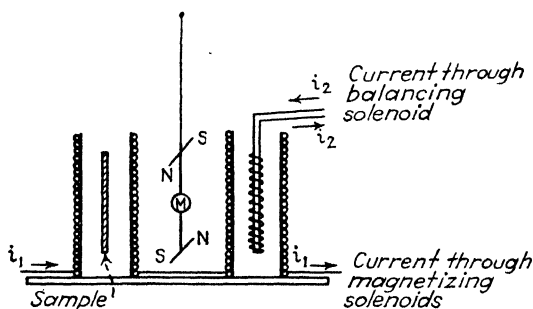



FIG. 11.7.—Bozorth magnetometer.

two methods illustrate the general principle that a substance with a permeability greater than unity tends to move into a region of strong field, and vice versa.

A magnetometer method of determining the premeability of a sample in the form of a long thin rod is illustrated in Fig. 11.7. An astatic magnetometer similar to the movement in the astatic galvanometer is placed midway between two similar solenoids which have their axes parallel to one another and to the axis of suspension of the small test magnets. The solenoids are wound in opposite senses and their positions so adjusted that there is no tendency for the movement to deflect when they are traversed by a current. However, if the sample of material is inserted in one solenoid, the magnetic moment induced by the field acts on both suspended magnets in a similar sense and the movement is subject to a torque. A compensating solenoid of the same shape as the sample is inserted in the second large solenoid and the current through it adjusted until the torque exerted by it on the movement exactly compensates that exerted by the sample. Under these circumstances the magnetic moment of the sample is equal to the calculable magnetic moment of the compensating solenoid. The actual magnetic moment per unit volume of the sample can be calculated approximately from this

total moment and a knowledge of the geometry. The current in the magnetizing solenoid determines **H** and from **H** and $\mathbf{m}_v$ the value of μ can be obtained. For a more detailed discussion of these methods and the others that have been devised for the measurement of permeability the reader is referred to treatises on magnetic measurements.[1]

11.5. Diamagnetism and Paramagnetism.—Experiments of the foregoing types show that some substances have values of μ greater than unity, and for others μ is less than unity. This is in distinction to the values found for the dielectric constants which are all greater than unity. If μ is less than 1, *i.e.*, if χ_m is negative, the substance is said to be *diamagnetic*, and if μ is greater than 1, *i.e.*, if χ_m is positive, the substance is said to be *paramagnetic*. The typical diamagnetic and paramagnetic substances have very small values of χ_m. The *ferromagnetic* substances that will be considered later, while not in general amenable to this simple treatment, can be considered as extreme examples of paramagnetic materials, for they exhibit values of μ of the order of 10^3. Table II lists the susceptibility in rational units of some of the more common diamagnetic substances.

TABLE II

Substance	Susceptibility, χ_m	Specific susceptibility, emu./gm.
Bismuth	-16.7×10^{-5}	-135×10^{-8}
Quartz	-1.51×10^{-5}	-49×10^{-8}
Water	-0.88×10^{-5}	-72×10^{-8}
Mercury	-3.23×10^{-5}	-19×10^{-8}
Silver	-2.64×10^{-5}	-20×10^{-8}
Lead	-1.69×10^{-5}	-12×10^{-8}
Copper	-0.94×10^{-5}	-8.6×10^{-8}
Argon (N.T.P.)	-0.945×10^{-8}	-45×10^{-8}
Hydrogen (N.T.P.)	-0.208×10^{-8}	-197.0×10^{-8}

The specific susceptibility is the susceptibility per unit mass rather than per unit volume, and these are given in electromagnetic units as this system is generally employed in the investigation of atomic magnetism.[2] To obtain the susceptibility per molecule these entries must be multiplied by the molecular weight and divided by Avogadro's number. Thus the atomic susceptibility for bismuth is -280×10^{-30} and for hydrogen -3.25×10^{-30} emu. To convert these to practical units they

[1] STONER, "Magnetism and Matter," Methuen & Co., Ltd., London, 1934; SPOONER, "Properties and Testing of Magnetic Materials," McGraw-Hill Book Company, Inc., New York, 1927.

[2] In these units the permeability μ' is $1 + 4\pi\chi'_m$ where χ'_m is the ordinary volume susceptibility in these units. The specific susceptibility is χ'_m divided by the density.

must be multiplied by the ratio of the emu. to the practical moment per unit field which is $(4\pi)^2 \times 10^{-13}$. Thus the moment developed per unit field for hydrogen in practical units is -5.2×10^{-41}.

The predicted diamagnetic susceptibility of an atomic medium can be calculated from first principles. If an electron is circulating in an orbit with an angular velocity ω_0 in the absence of a field, the atomic centripetal force on it must be $-m\omega_0^2 r$ to balance the centrifugal tendency. If a field **H** is applied in the direction of rotation, an additional centrifugal force $e\omega\mu_0 Hr$ makes its appearance where e is the electronic charge and ω is the angular velocity in the presence of the field. If the atomic centripetal force remains the same, the centrifugal forces in the absence and presence of the field may be equated, yielding $m(\omega^2 - \omega_0^2) = e\omega\mu_0\mathbf{H}$. Or, since the change in angular velocity is small for any fields that can be achieved in the laboratory, the increase in angular velocity, $(\omega - \omega_0)$, can be written $\delta\omega$ and $(\omega + \omega_0)$ as 2ω or

$$\delta\omega = \mu_0 e/2m\ \mathbf{H}$$

Since the magnetic moment **m** is proportional to $-\omega$, the induced magnetic moment is $\delta\mathbf{m} = -\dfrac{\mathbf{m}\mu_0 e}{2\omega m}\mathbf{H}$. As $\mathbf{m}/\omega$ can be written $\mathbf{m}I/\mathbf{p}$, where I is the effective moment of inertia and the characteristic ratio of $\mathbf{m}/\mathbf{p}$ is $\mu_0 e/2m$ for a circulating electron, the change in magnetic moment induced by the field becomes

$$\mathbf{m}_i = -\left(\frac{\mu_0 e}{2m}\right)^2 I\mathbf{H} \tag{11.12}$$

The effective moment of inertia for an atom is $m\sum_{j=1}^{j=n}(\bar{x}_j^2 + \bar{y}_j^2)$, where m is the electron mass and the quantities in the bracket are the mean-square orbital coordinates of the jth electron. The summation is over all the electrons. Assuming complete spherical symmetry, the radius vector to the jth electron is as likely to be pointing in one direction as another. Under these conditions by the Pythagorean theorem

$$\bar{x}_j^2 = \bar{y}_j^2 = \bar{z}_j^2 = \left(\frac{\bar{r}_j^2}{3}\right) \quad \text{or} \quad I = \frac{2m}{3}\sum_{j=1}^{j=n}\bar{r}_j^2$$

The coefficient of **H** in Eq. (11.12) with this value of I is the induced atomic magnetic moment per unit field. It is negative, independent of the sign of e, since by Lenz's law the induced field must be in such a sense as to oppose the inducing one. The equation also shows that the induced moment is proportional to the product of the number of electrons in the atom and the mean orbital area. This has been verified for quite a number of atomic types and is born out by the preceding table. Also if the numerical values of μ_0, e, and m are substituted in the coefficient of **H** in Eq. (11.12) and the result equated to the atomic susceptibility of hydrogen given previously it is found that the mean value of the linear dimensions of the electron orbit in hydrogen is of the order of 10^{-10} m., which is known to be approximately correct from other lines of evidence. From these results it is evident that ordinary electromagnetic theory is adequate to account for diamagnetic phenomena.

In connection with the Stern-Gerlach experiment it was mentioned that certain atoms are found to have natural magnetic moments asso-

ciated with them owing to unbalanced electron spins or orbits. When a magnetic field is applied to such an atom, the axis of the magnetic moment is induced to precess about the direction of the field, yielding a net positive contribution to the total flux. Of course, the diamagnetic tendency is present also, but the magnitude of the natural magnetic moment is such that it more than compensates for the change in orbital rotation of the electrons. Table III gives the susceptibility and specific susceptibility in emu. per gram for some of the more common paramagnetic substances.

TABLE III

Element	χ_m	Specific susceptibility, emu./gm.
Liquid oxygen	3.46×10^{-3}	$31,000 \times 10^{-8}$
Palladium	8.25×10^{-4}	540×10^{-8}
Platinum	2.93×10^{-4}	110×10^{-8} (20°C.)
Aluminum	2.14×10^{-5}	63.0×10^{-8} (20°C.)
Oxygen	1.79×10^{-6} (N.T.P.)	106.2×10^{-6} (20°C.)
Air	3.65×10^{-7} (N.T.P.)	24.2×10^{-6} (20°C.)

It will be noted that these are larger in magnitude than diamagnetic susceptibilities but still very small compared to unity. The paramagnetic susceptibility of a substance is a function of the temperature. The reason for this is that the thermal motion of the molecules tends to annul the net orientation in the direction of the field.

The variation of paramagnetic susceptibility with temperature is somewhat similar to the dependence of the dielectric constant of a polar gas on temperature. The diamagnetic tendency is independent of temperature and is analogous to the electrostatic distortion polarization. The average effective component of the magnetic moment of a molecule in the direction of the field, when subject to random thermal agitation, can be calculated in the same way as the analogous electrostatic quantity in Sec. 3.1. However, the analysis holds only above a certain temperature, for as the thermal agitation becomes less, the interatomic forces tend to make large groups of molecules act together as a unit and ferromagnetic phenomena make their appearance. If θ is written for this lower temperature limit (called the *Curie temperature*), the mean effective atomic moment would be written

$$\overline{\mathbf{m}} = \frac{\mathbf{m}^2}{3k(T - \theta)}\mathbf{H}$$

Here k is Boltzmann's constant in joules per degree per molecule, T is the

absolute temperature, and θ is the characteristic temperature of the substance. The behavior of the substance is linear only for values of $T >> \theta$. For values of T in the neighborhood of θ or for large magnetic fields the magnetic energy is large in comparison with the thermal energy and the alignment of the moments with the field tends to become complete. This phenomenon is known as *saturation*. It occurs only at very low temperatures for ordinary substances and magnetic fields that can be produced in the laboratory.

In addition to ordinary paramagnetic substances with values of μ very close to unity, there are certain ions as salts or in solution that show very strong paramagnetism. They have values of μ midway between unity and the large values characteristic of ferromagnetism. These substances fall into three groups. There are the rare-earth ions from La to Cp with values of μ lying between 1.5 for Sm to 10.5 for Ds and Ho; the ferric group of ions from V to Cu with values of μ from 1.75 for V and Cu to 6 for ferric iron salts; and the group from W to Pd with in general somewhat smaller permeabilities. These large permeabilities or magnetic moments are characteristic of the ions and can be accounted for adequately on the quantum theory of the resulting electron configuration.

11.6. Ferromagnetism.—The most important magnetic materials for practical purposes are the ferromagnetic ones such as iron and its alloys. These are characterized by very large values of $\mathbf{m}_v$. Equation (11.2), of course, applies for these materials, but Eq. (11.8) must be considered as defining a variable quantity μ known as the permeability not as implying a simple linear relation between **B** and **H**. The approximation to which μ can be considered as a constant is in general a poor one for iron and its alloys. The ratios of the induction or magnetic moment per unit volume to the effective field are found to depend on many factors which can be known only if the previous mechanical, thermal, and magnetic history of the sample is given. Thus in general neither **B** nor $\mathbf{m}_v$ can be considered as simply functions of **H**. When the previous history of the sample has been established in such a way as to ensure a certain regularity of behavior, as by the frequent reversal of an applied field of constant magnitude, a functional relation between **B** and **H** is established, but even in this case the function is double-valued and the appropriate value of **B** depends on whether **H** is increasing or decreasing. The actual phenomena including the effects of thermal and mechanical factors are very complex and only the general principles underlying them are understood. Recent investigations have contributed a great deal to our knowledge of the atomic and microscopic processes that are involved and it is more instructive to consider the subject first from this point of view. It will be possible here to give only a brief general

outline of our present knowledge of the phenomena; for a more complete account the special treatises and current articles must be consulted.[1]

It has been seen in the preceding sections that the speeding up of unfavorable electron orbits and the slowing down of favorable ones, which is induced by the application of a magnetic field, results in a general fundamental diamagnetic effect in all materials. However, if permanent moments are associated with the atoms by reason of unbalanced electron spins or orbital motions, these more than compensate the diamagnetic tendency and a net paramagnetic effect is observed. Certain ions were mentioned at the conclusion of the preceding section that have uncompensated spins owing to so-called *internal electrons* which are closely bound to the atom and remain associated with the ion in a crystal. The three groups of elements in which this situation occurs display particularly strong paramagnetism. It is among these substances that ferromagnetism is observed. That it is a phenomenon of the electron spin is brought out by the fact that the ratio of magnetic moment to angular momentum for ferromagnetic substances is approximately that characteristic of the spinning electron (Sec. 10.7). For true ferromagnetism, however, another condition must be fulfilled and this is associated with the interatomic forces that hold the ions in their position in a crystal lattice. This is borne out by the fact that ferromagnetism is observed only in crystalline forms. The so-called *exchange forces* which are largely responsible for crystal phenomena have no large-scale analogues but may be considered to arise through the identical nature of the electrons in the crystal. In the close proximity of a crystal lattice one electron is essentially shared by a number of ions, and this gives rise to forces of interaction. In the neighborhood of a critical ratio of spacing to ion diameter these forces appear to favor the alignment of all the unbalanced internal-electron spins of the ions in a certain microscopic crystal domain. These forces are very strong, corresponding to applied fields of the order of 10^7 oersteds or amp.-turns per centimeter. They are effective over a domain about 2.5×10^{-3} cm. on a side. This includes approximately 10^{15} atoms. The effect of the exchange forces is to make all these atoms act together more or less as a single atom. This point of view is born out by many lines of evidence. One of these comes from the microscopic examination of the surface of a magnetized crystal. Etching and other methods of testing the surface show characteristic microscopic patches with uniform magnetic properties. Also

[1] BITTER, "Introduction to Ferromagnetism," McGraw-Hill Book Company, Inc., New York, 1937; MESSKIN and KUSSMANN, "Properties of Ferromagnetic Alloys," Verlag Julius Springer, Berlin, 1932; ELLIS and SCHUMACHER, *Bell System Tech. J.*, **14**, 8 (1935); BOZORTH, *Bell System Tech. J.*, **15**, 63 (1936); ELMEN, *Bell System Tech. J.*, **15**, 113 (1936); VAN VLECK, *Rev. Mod. Phys.*, **17**, 27 (1945); BOZORTH, *Rev. Mod. Phys.*, **19**, 29 (1947).

magnetization in certain field-strength regions is not a continuous process but takes place in a series of infinitesimal discrete steps. This can be shown by inserting the sample in a coil connected to an amplifier and loud-speaker. If the field surrounding the sample is slowly increased, individual clicks are heard in the speaker which represent small discrete flux increments. This is known as the *Barkhausen effect.* The steplike nature of the magnetization curve due to the orientation of these crystal domains in this region is indicated in Fig. 11.8.

FIG. 11.8.—Magnification of the central region of a magnetization curve to show its steplike character due to the discontinuous nature of the magnetizing process (Barkhausen effect).

Since ferromagnetism is a crystal phenomenon, it is most unambiguously studied in large single crystals where the directional effects can be associated with the crystal axes. Experiments of this type show that magnetization is much more readily produced along certain crystal axes than along others. In the case of iron a small field is sufficient to line up all spins in the so-called 1-0-0 direction, which means in the direction of one of the natural Cartesian axes of the body-centered cubic iron crystal. A larger field is necessary to obtain complete saturation (alignment of all the uncompensated electron spins) along either the face or body diagonals.

FIG. 11.9.—(*a*) Crystal directions and relative ease of magnetization in the body-centered iron crystal. (*b*) Magnetization curves for a single iron crystal for the three crystal directions.

This is illustrated in Fig. 11.9. Thus there are six directions (+ and − for the three axes) in which it is relatively easy to magnetize the crystal, all other directions being more difficult. Ordinary samples are, of course, composed of a large number of small crystals arranged in random orientations. When a magnetic field is applied to such an aggregate, the early stages of magnetization correspond to the sudden orientation of domain

moments into those directions of easy magnetization which have a component in the direction of the magnetizing force. This produces the Barkhausen effect in the central region of the magnetization curve. As the magnetizing force increases, these domain moments are gradually oriented against the natural crystal forces till they approach the direction of the field itself. This corresponds to the farther region of the magnetization curve associated with saturation.

There are also volume changes associated with magnetization. At small values of **H** those domains that are being magnetized in a favorable direction appear to increase in volume with respect to those unfavorably oriented. This accounts for various phenomena observed in connection with the region of the magnetization curve near the origin. The change in relative dimensions of a sample on magnetization is known as *magnetostriction*. It is important for certain practical purposes and also plays

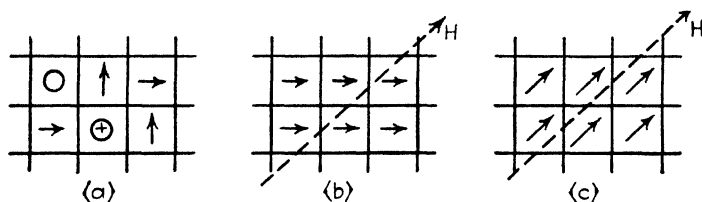



FIG. 11.10.—Stages in ferromagnetization. (a) Unmagnetized randomly oriented domains. (b) Domains magnetized in the easy direction. (c) Domains oriented in the direction of the field.

an essential rôle in the fundamental phenomena of magnetization. In the case of iron the dimensions tend to increase in the direction of **H** and decrease perpendicular to **H**; for nickel the opposite is true. As these changes are resisted by the crystal forces, they resist magnetization to this extent. The work done against the crystal forces is not all conserved and part appears in the form of heat which raises the temperature of the sample. An externally applied strain changes both the magnitude of the magnetic effects and their directional properties. Local random strains, such as are produced in a lattice by the presence of impurities or by certain heat treatments, have a profound magnetic effect. A substance such as steel can be rendered very hard by rapid cooling and the crystal strains introduced in this way also render the orientation of magnetic domains much more difficult. Samples of this character are also said to be magnetically "hard," for very large fields must be employed to alter their magnetization, and they can withstand considerable mechanical shock without change in magnetic moment. Thus the previous thermal history of the sample is seen to be of great importance. In fact, the majority of the progress in the development of useful magnetic materials has resulted from the choice of a suitable alloy and the careful control of its heat treatment. Samples of permalloy

cooled in a magnetic field have yielded values of μ as high as 10^6 and samples of Alnico 5 (51 per cent Fe, 24 per cent Co, 14 per cent Ni, 8 per cent Al, 3 per cent Cu) once magnetized are capable of retaining an induction of the order of 1.25 webers per square meter permanently. Of course, at very high temperatures the thermal agitation and expansion destroys the critical ionic spacing and with it ferromagnetic effects. This generally occurs for iron in the range from 500° to 700°C. (Curie temperature).

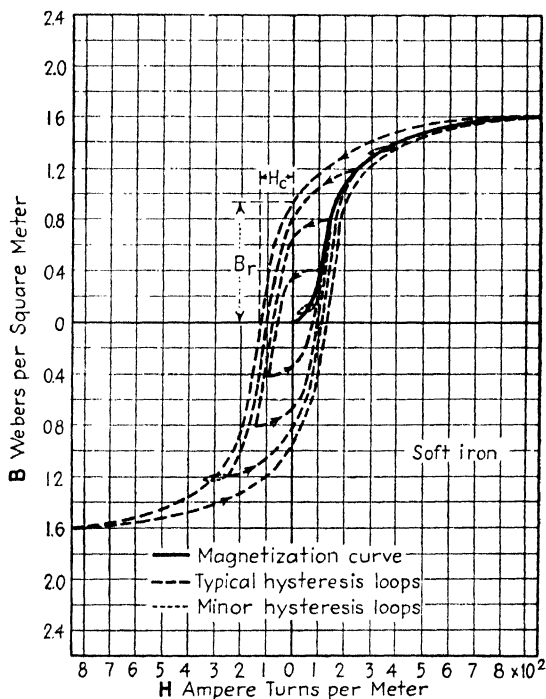



FIG. 11.11.—Normal magnetization and representative hysteresis curve.

11.7. Hysteresis Curves and General Magnetic Properties.—The phenomena that are of particular interest for alternating-current work are those which appear when the sample is subjected to a magnetic field that is periodically reversed in direction. After a few reversals of this type, the sample enters a cyclic condition in which the induction is determined by the magnetizing field and its sense of alteration. The closed curve that gives the relation between B and H is known as a *hysteresis curve.* The particular curve depends on the maximum value of the applied field. A normal hysteresis curve is one that is obtained by varying the applied field between positive and negative values of equal magnitude. A family of these is shown in Fig. 11.11. The solid curve traced out by the sharp terminations of the hysteresis curves

obtained for various values of the maximum field is known as the *normal magnetization curve.* It differs somewhat in the neighborhood of the origin from the ordinary magnetization curve which gives the relation between B and H for increasing values of H, starting from an initially demagnetized sample. Demagnetization is accomplished by putting the sample through successively smaller normal hysteresis loops until the maximum value of the alternating field has been reduced to zero. The intersection of the normal hysteresis loop with the ordinate axis is the *remanent induction* B_r which is the residual flux density retained by the sample when the field is reduced to zero after having achieved a particular maximum value. The intersection of the loop with the abscissa axis is called the *coercive force* H_c and represents the field that must be applied in the reverse direction to reduce the flux through the sample to zero. If a large constant field and a small alternating one are both applied to the sample, the relation between B and H is known as a *minor hysteresis loop.* Since the entire diagram can be filled with hysteresis curves, it is evident that little significance can be attached to any constant value of μ. But the *permeability* or *normal permeability* is defined as the ratio of B to $\mu_0 H$ given by the normal magnetization curve. It is, of course, a function of H. The *differential permeability* is used to refer to the slope of this curve, dB/dH divided by μ_0. This is proportional to the instantaneous inductance of an iron-core coil, for this latter is proportional to the rate of change of flux with respect to current. Hence the inductance of such a coil will in general vary over a wide range. Another quantity that is frequently used is the *incremental permeability* associated with a minor hysteresis loop. It is the slope of the line joining the tips of the small loop divided by μ_0. It represents the effective permeability for a small alternating current superimposed on a large direct current. It is seen from Fig. 11.11 that the slope of this line is smaller than that of a normal hysteresis curve through the point. The *effective alternating-current permeability* is the average value of the permeability associated with a normal hysteresis curve. It is determined most simply by measuring the effective impedance presented by an iron-core coil to an alternating current. It is, of course, a function of the maximum applied magnetic field.

In order to obtain curves of the type of Fig. 11.11 the sample must be of such a shape that it is possible to measure both the magnetic field applied to the specimen and also the resulting magnetic induction. The simplest geometrical form for calculating the field is the toroidal ring. Samples in the form of straight rods can also be tested with special equipment,[1] but the general principles involved are illustrated by a consideration of the torus. The experimental setup is shown schemati-

[1] SPOONER, *op. cit.*

cally in Fig. 11.12. The upper circuit controls the magnetizing current through the winding on the toroidal specimen and the ballistic galvanometer measures the changes in flux, from which the changes in induction can be calculated. The resistance R_s is used to vary the sensitivity of the instrument and the damping key K controls its motion. The lower standard solenoid circuit is for standardization so that the galvanometer deflections can be interpreted in terms of flux increments. The galvanometer must be recalibrated after making any change in sensitivity, *i.e.*, any alteration in R_s. The procedure for obtaining a hysteresis loop is as follows. S_1 and S_3 are closed and the magnetizing current, i_1, adjusted to the desired maximum value for the loop by means of R_1. Then 20 or 30 reversals of the current are made at intervals of a few seconds by means of S_1 until the specimen has reached the characteristic cyclic state. This may be judged by the uniformity of the galvanometer throws. Once achieved this state must be maintained throughout the loop, *i.e.*, the current can be reversed in direction only after its maximum value has been reached in the appropriate sense. R_s is then adjusted till the galvanometer gives approximately a full-scale deflection upon reversal. S_1 is then opened and S_2 closed and the solenoid current i_2 adjusted by means of R_2 until a reversal of S_2 gives a comparable throw t_2. The flux linkage with m_2 is $\mu_0 A_2 i_2 n_2' m_2$; hence the change in flux linkage on reversal of S_2 is $2\mu_0 A_2 i_2 n_2' m_2$ and this is proportional to t_2. If t is the throw obtained on altering the flux through the ring, δB is given by

FIG. 11.12.—Apparatus for obtaining the magnetic characteristics of a ring specimen.

$$\delta B = \frac{2\mu_0 A_2 i_2 n_2' m_2}{t_2 A_1 m_1} t$$

since the constant of proportionality is the same and $\delta\phi = A_1 \delta B$. All the quantities in this coefficient are known. The magnetizing field in the ring is also readily obtained. If the radius of cross section is small in comparison with the ring radius, H is approximately uniform and equal to $i_1 n_1'$, hence

$$\delta H = n_1' \delta i_1$$

The reversals of S_1 with S_3 closed locate the points a and b of Fig. 11.13. Other points on the loop are obtained by manipulating the switches S_1 and S_3 in the proper sequence. R_3 is first set at such a value that the opening of S_3 will cause a change in i_1 by about 10 per cent, S_1 being, say, in its upper position and S_3 closed. S_3 is then opened and

δi_1 and t determine the values of δH and δB associated with the passage of the specimen from a to c. S_1 is then opened and the point d located in the same manner. S_1 is closed in its lower position locating e and finally S_3 is closed reaching b. The cycle is then repeated in the reverse direction passing through c', d', and e'. The value of R_3 is then altered and another set of points obtained. Care must be exercised to preserve the proper sequence or the specimen will depart from the particular loop being investigated. Special switches are available which automatically perform this sequence of operations. The normal magnetization curve can be obtained without the use of R_3 and S_3 by putting the specimen in a series of cyclic states obtained by varying R_1 and recording only the values of i_1 and t that determine the positions of the tips of the hysteresis loops. The normal magnetization curve is the locus of these tips. Minor loops are obtained by inserting a constant emf. directly in series with the magnetizing winding on the ring. These procedures are somewhat tedious and for rapid qualitative work circuits can be devised very simply for rendering hysteresis loops visible on the screen of an oscillograph. For instance, an alternating current is sent through the magnetizing winding and the potential for one set of plates obtained from the drop in a series resistance. This deflection is then proportional to H. If an inductive reactance large in comparison with m_1 and the circuit resistance is placed in series with m_1, the current in this circuit is proportional to the flux through the ring; hence, if the drop across a small series resistance is applied to the other pair of oscillograph plates, the deflection along this axis will be proportional to B.

FIG. 11.13.—Point-by-point determination of a hysteresis loop.

One important application of the hysteresis curve is in connection with the energy loss associated with the traversal of a loop by the substance. This may be considered from the point of view of the rate at which the external circuit performs work. By definition this is equal to $\mathcal{E}i$, where $\mathcal{E}$ is the emf. across the terminals of the magnetizing coil and i is the current through it. Neglecting resistance $\mathcal{E} = -n_1 A_1 \frac{\partial B}{\partial t}$ and $\mathcal{H} = n_1 i = lH$, where l is the length of the specimen. The work per cycle is thus

$$W = \int \mathcal{E}i\,dt = -V\int H\frac{\partial B}{\partial t}\,dt = V\oint H\,dB$$

where V is written for the volume Al of the specimen. Thus the integral of $H\,dB$ around the loop, which is equal to its area, is equal to the energy dissipation per unit volume per cycle. Using the scale of Fig. 11.11, the area is in terms of joules per cubic meter or in units of 10 ergs per cubic centimeter. This result could also be achieved by recalling that the change in energy per unit volume due to a change in moment can be written $H\,dm_v$. From Eq. (11.2) $dm_v = dB - \mu_0\,dH$ and since the integral of $H\,dH$ around any closed path is zero, the change in magnetic energy due to the traversal of a closed path can be written $\oint H\,dB$ in agreement with the previous derivation. The area of the hysteresis loop is of great importance in alternating-current machinery since it represents the loss of electrical energy or the generation of heat per cycle. Since the shapes of the hysteresis loops vary widely for different magnetic materials it is difficult to obtain an expression for W/V as a function of H_{max} or B_{max}. However, an approximate expression which is useful for comparing the relative merits of different materials for alternating-current work is due to Steinmetz. It is $W/V = \eta B_{max}^{1.6}$. The constant η is known as the *Steinmetz coefficient* for the material and in terms of gauss and ergs per cubic centimeter it has values ranging from 0.0002 for permalloy through 0.005 for soft steel to 0.075 for glass-hard tool steel.

11.8. Special Ferromagnetic Materials.—A very large number of ferromagnetic alloys have been developed for various purposes, but it is possible here to mention only a few of the more interesting or important types. They fall more or less into two classes, depending on whether they are mechanically and magnetically soft or hard. Magnetically soft materials are required for alternating-current machinery and for power and communication transformers. These devices will be discussed in more detail in a subsequent chapter, but it may here be mentioned that the requirements on the material for the core of a communication transformer are very stringent. For the true reproduction of an imposed wave form the characteristics should be linear; this implies a hysteresis loop with as slight a curvature and as small an area as possible. The permeability should be high and constant over a wide range of magnetization. Furthermore, the electrical resistance should be high in order to minimize the eddy current losses caused by currents induced in the magnetic material. From Sec. 10.2 it is evident that the component of the eddy currents in phase with the inducing emf. is proportional to the conductivity. Thus, if the conductivity is small, which means that the resistivity is large, the joule losses in the material are small. For direct-current and permanent magnet applications, however, the linearity and resistivity are of little importance. A high permeability is the chief desideratum for these purposes and for permanent magnets the remanent induction and coercive force should be as large as possible. The product of these latter quantities can be taken as a measure of the quality of a permanent-magnet material. It is evident that substances of this type would be classed as magnetically hard. Curves giving the magnetization per unit volume as a function of the magnetizing force for various representative materials are shown in Fig. 11.14. Frequently the permeability $B/\mu_0 H$ is plotted as a function of B, and three curves of this type are shown in Fig. 11.15.[1]

[1] For further detailed material of this type see the references previously cited,

Pure iron is itself one of the most important magnetically soft materials. Minute quantities of impurities greatly affect its magnetic characteristics; the principal detrimental ones are carbon, oxygen, nitrogen, and sulphur. Purification and careful heat treatment in an atmosphere of hydrogen can produce permeabilities as high as 3.4×10^5. The coercive force in low and these high permeabilities are reached for

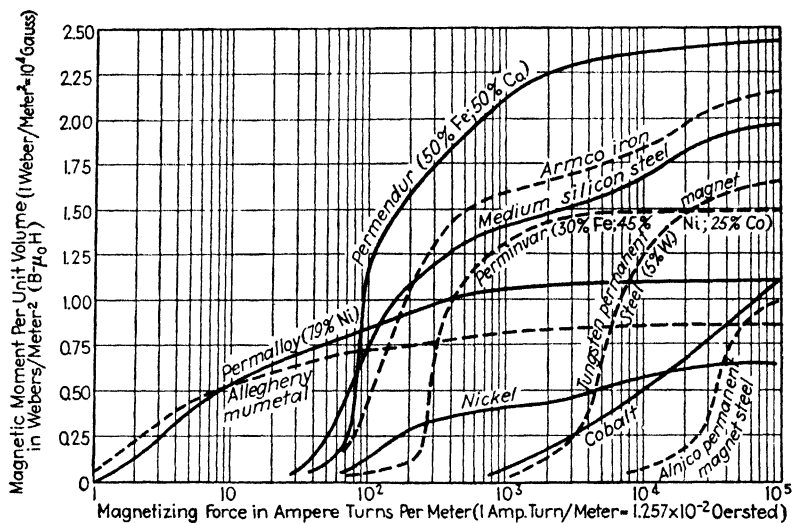



FIG. 11.14.—Magnetic curves of the more useful ferromagnetic elements and alloys.

low magnetizing forces. The area of a hysteresis loop is small, corresponding to a loss of the order of 200 ergs per cubic centimeter cycle at 14,000 gauss which can be compared with about 3,000 ergs per cubic centimeter cycle at the same maximum flux density for ordinary iron. Probably the most commonly employed magnetic material

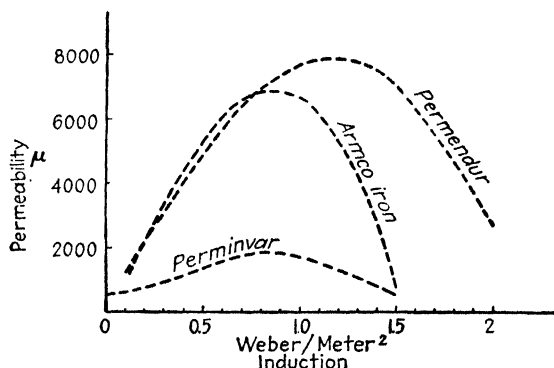



FIG. 11.15.—Permeability curves.

is *silicon steel.* The addition of silicon in small quantities (up to about 4 per cent) by its deoxidizing action and tendency to reduce hard carbon results in a comparatively soft magnetic material that has a high resistivity and can be produced relatively

technical handbooks, and Macmillan, *Gen. Elec. Rev.*, **39**, 225, 282 (1936); see also *Electronics*, May, 1936, p. 30.

cheaply. Alloys containing about 80 per cent iron and 20 per cent nickel, known as *permalloys*, show very high permeabilities at low flux densities. Also their hysteresis losses are small. The addition of a few per cent of chromium or molybdenum increases the initial permeability and the resistivity as well as rendering the heat treatment less critical. Other alloys containing approximately 50 per cent iron and 50 per cent nickel, the *hypernik* and *nicaloi* group, are only slightly inferior in magnetic properties and are more magnetically rugged. The addition to these of small amounts of manganese or copper produces useful alloys with the same general properties such as *mumetal*. All these substances are widely used in the communication industry. Other substances are characterized particularly by the constancy of their permeability. Iron or permalloy dust, when pressed together with a resin binder, forms a material having a very constant permeability and a high resistance. Dust-core coils are widely used for inductances at high frequencies. Certain alloys (in the neighborhood of 55 per cent Ni, 37 per cent Fe, 8 per cent Cu), which are known as *isoperms*, have approximately the same characteristics. In the region of low flux

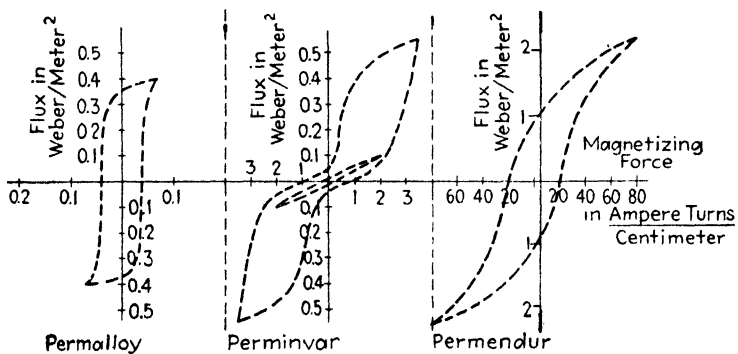



FIG. 11.16.—Typical hysteresis loops of various alloys.

densities the *perminvars*, which are alloys containing about 45 per cent Ni, 30 per cent Fe, and 25 per cent Co, subjected to special heat treatment, display particularly constant permeabilities and associated linearity and low losses, as indicated in Fig. 11.16. However, these substances are unstable magnetically, for at high flux densities (above about 1,000 gauss) the hysteresis loop completely changes its character and once the substance has entered this region, its desirable characteristics are lost and can be regained only by a subsequent heat treatment.

Many valuable magnetically hard alloys have also been developed. *Permendur* is an alloy of half iron and half cobalt, with sometimes the addition of a few per cent of vanadium to increase its resistivity. It has the highest known saturation magnetization, reaching approximately 25,000 gauss at 1,000 amp.-turns per centimeter. It is thus particularly valuable for direct-current electromagnets. It also retains a high permeability for alternating currents in the presence of a large constant flux which makes it a very useful material for the construction of inductances which must retain their reactance in the presence of a large direct-current component. For a permanent-magnet material the retained flux density or remanence must be high as must also be the coercive force. The free poles that characterize a useful permanent magnet tend to demagnetize the specimen and this tendency must be resisted by the coercive force. Hard chrome-tungsten and -molybdenum steels have remanent inductions of the order of 10,000 gauss and coercive forces of about 60 amp.-turns

per centimeter and make very satisfactory permanent magnets. *Alnico* has an even greater remanent induction and a coercive force about eight times as great. It also has desirable thermal properties and is light in weight, but it is brittle and must be ground after being cast.[1]

There are also iron alloys which are practically nonmagnetic. One of these, which is composed of 88 per cent Fe and 12 per cent Mn, shows very feeble magnetization and a practically zero hysteresis loss. Another is composed of 68 per cent Fe and 32 per cent Ni and has magnetic properties comparable with typical paramagnetic substances. Conversely a substance which contains neither iron, cobalt, nor nickel and which is known as *Heusler's alloy* (61.5 per cent Cu, 23.5 per cent Mn, 15 per cent Al) displays strong magnetic properties. It has a remanent induction of about 2,500 gauss, a coercive force of about 6 amp.-turns per centimeter, and a Steinmetz coefficient of about 0.003. Alloys have also been developed for various purposes involving magnetostriction. As an instance oscillators can be stabilized at the natural period of mechanical vibration of such a material. An alloy containing 54 per cent Fe, 36 per cent Ni, and 10 per cent Cr shows a large magnetostrictive effect and small current and temperature coefficients of the natural frequency $\left(\frac{1}{\nu}\frac{\partial \nu}{\partial T} = 3 \times 10^{-6}\text{ per degree Centigrade, } \frac{1}{\nu}\frac{\partial \nu}{\partial H} = 10^{-6}\text{ per oersted}\right)$. Certain iron-nickel alloys, when properly tempered, show a negligible change in elastic properties with temperature and are known as *elinvars*. *Monel metal* and another alloy consisting of about 70 per cent Fe and 30 per cent Ni show a large temperature coefficient of magnetization at room temperature; they are useful for the thermal control of the magnetic properties of a circuit.

11.9. The Magnetic Circuit.—The determination of the magnetic induction at every point in the presence of irregularly shaped masses of magnetic material is a boundary-value problem that is very difficult to solve rigorously. However, the concepts of amperian currents or pole distributions can be used to obtain a general description of the field or induction in the various regions. In the case of a simple substance in the form of a sheet placed normally in a region of induction **B** the continuity of the normal component of **B** shows that **B** is the same inside and outside of the sheet. Since **B** and **H** are proportional to one another in a simple substance ($\mathbf{B} = \mu\mu_0\mathbf{H}$), the technique of Sec. 2.5 can be taken over directly. Excluding regions actually containing currents, a scalar potential satisfying Laplace's equation exists, Ω, of Sec. 9.4, and **H** is its negative gradient. Thus **H** is the analogue of **E** and **B** of **D** in Sec. 2.5. Likewise μ replaces κ and μ_0 replaces κ_0. The boundary conditions div $\mathbf{B} = 0$ and **curl** $\mathbf{H} = 0$ are of the same form as in the electrostatic case. Thus, in the case of a sphere of a simple magnetic substance of permeability μ and radius a placed in free space in which the field had previously been uniform throughout of value $\mathbf{H}_0$, the potentials within and without the sphere are given by

[1] For further information on the subject of permanent-magnet steels see EVERSHED, *J. Inst. Elec. Eng.*, **58**, 780 (1920); SCOTT, *Am. Inst. Elec. Eng. Trans.*, **51**, 472 (1932); EDGAR, *Gen. Elec. Rev.* **38**, 466 (1935); WILLIAMS, *Elec. Eng.*, **55**, 19 (1936).

$$\Omega_i = -\left(\frac{3}{\mu + 2}\right)H_0 r \cos\theta \qquad \text{(inside)}$$

$$\Omega_0 = -\left[1 - \left(\frac{\mu - 1}{\mu + 2}\right)\frac{a^3}{r^3}\right]H_0 r \cos\theta \qquad \text{(outside)}$$

Thus the field inside is $-\frac{\partial\Omega}{\partial x}$ $(x = r\cos\theta)$ or $\frac{3\,\mathbf{H}_0}{\mu + 2}$ or the induction inside is $\frac{3\mu\,\mathbf{B}_0}{\mu + 2}$, where $\mathbf{B}_0$ is the induction that previously existed or the value at a great distance. The field outside is that which previously existed (from the first term) plus that which would be produced by a current vortex of moment $\mathbf{m} = 4\pi\mu_0\left(\frac{\mu - 1}{\mu + 2}\right)a^3\mathbf{H}_0$. This corresponds to a uniform moment per unit volume of $3\left(\frac{\mu - 1}{\mu + 2}\right)\mu_0 H_0$, which is in agreement with the relation $\mathbf{m}_v = (\mu - 1)\mu_0\mathbf{H}$, where $\mathbf{H}$ is the field inside the sphere. A cylinder of a simple magnetic substance with its axis normal to the direction of the field can be handled in an analogous way. The other extreme calculable case is that of a cylinder very long in comparison with its diameter placed with its axis parallel to an external field $\mathbf{H}_0$. From the continuity of the tangential component of $\mathbf{H}$ the field is the same inside and outside and the internal induction is

$$\mu\mu_0 H_0 = \mu B_0$$

Another analogy, of a different type from that between charges and poles, which is very useful for the solution of many magnetic problems, is that existing between induction and current density. Both these vectors are solenoidal, *i.e.*, their divergence vanishes. By comparing the equations

$$\mathbf{i}_v = \sigma\mathbf{E} \qquad \text{and} \qquad \mathbf{B} = \mu\mu_0\mathbf{H}$$

it is seen that $\mu\mu_0$ is the analogue of the conductivity σ. Furthermore, in the absence of charges in the one case and currents in the other for the region under consideration all the vectors satisfy Laplace's equation if μ and σ are constants. The lines of flow and induction are identical curves for analogous conduction and magnetic problems. Of course, the general current-flow problem is no easier to solve than the magnetic one. However, in electrical work one is frequently concerned with the situation in which the lines of flow are largely confined by and approximately parallel to the boundary of a limited circuital region. If the lines of flow are also uniformly distributed over the cross section of the circuit, it develops that there is a significant circuit parameter known as the resistance, R, which is equal to $l/\sigma A$, where l is the length of the

circuit and A its cross-sectional area. From the previous discussion it is evident that under the same conditions for the lines of induction an analogous quantity known as the *reluctance*, $\mathcal{R}$, will have the same significance for what can be called a magnetic circuit as R has for an electric one. The formal equations are

$$\mathcal{E} = Ri \qquad \mathcal{H} = \mathcal{R}\phi$$

$$\mathcal{E} = \oint E\, dl \qquad R = \frac{l}{\sigma A} \qquad \mathcal{H} = \oint H\, dl \qquad \mathcal{R} = \frac{l}{\mu\mu_0 A} \tag{11.13}$$

The approximations for the magnetic problem are somewhat less satisfactory than for the electric one. There is generally less difference between μ for the circuit and the surrounding medium than there is for σ, and this results in the circuit being less well defined. Also, σ is very closely a constant for good conductors, whereas it has been seen that μ is only very approximately so for the important ferromagnetic materials, and these are the only ones which have a large enough value of μ to make the analogy of practical importance. Furthermore, the ratio of linear to cross-sectional dimensions is generally much smaller for magnetic circuits and this increases the error introduced in assuming the lines of induction parallel to the boundaries which is involved in taking A as the mean area and l as the mean length of the circuit. However, in spite of these limitations the magnetic-circuit concept is of considerable practical importance. It should be pointed out that $\mathcal{H}\phi$ does not represent a rate of dissipation of energy as $\mathcal{E}i$ does in the electrical case. To the magnetic circuit linear approximation $\frac{1}{2}\mathcal{H}\phi$ represents the energy stored in the circuit.

FIG. 11.17.—Illustration of the magnetic circuit.

From the similarity of the equations it is evident that series and parallel magnetic circuits can be handled exactly as electric ones. The effective reluctance of a number of parallel paths is equal to the reciprocal of the sum of the reciprocals of the reluctances for the separate paths, and the reluctance of paths in series is equal to the sum of the reluctances of the paths making up the circuit. Figure 11.17 illustrates a simple electromagnet with an airgap. The magnetomotive force $\mathcal{H}$ is equal to ni and the reluctance of the circuit is $\dfrac{\left(\frac{L}{\mu} + l\right)}{\mu_0 A}$. This assumes that the effective area of the gap is the same as that of the iron circuit which neglects the spreading of the lines of induction in that region and is a satisfactory approximation only if $l^2 << A$. The total flux is given by

$$\phi = \frac{ni\mu_0 A}{l + \frac{L}{\mu}}$$

The magnetomotive force across the gap is

$$\mathcal{H}_g = \phi \mathcal{R}_g = \frac{ni}{1 + \frac{L}{l\mu}}$$

If L/l is of the order of 50 and μ is of the order of 5,000 all but about 1 per cent of the applied magnetomotive force appears across the gap. In one method of obtaining the magnetic characteristics of a straight-rod sample it is placed across the limbs of a heavy U-shaped yoke of high permeability which provides a low-reluctance path for the flux traversing the specimen. If the magnetomotive force due to a winding on the sample is $\mathcal{H}$, that appearing across the sample is by the previous argument approximately

$$\mathcal{H}_s = \mathcal{H}\left(1 - \frac{l_y \mu_s A_s}{l_s \mu_y A_y}\right)$$

where the subscript s refers to the specimen and subscript y to the yoke, if the fraction in the parentheses is small. The second term can generally be made a correction factor of only a few per cent, but it must be taken into account to obtain the magnetomotive force applied to the specimen. Other corrections may also be necessary to allow for the approximations involved in the magnetic-circuit concept.

11.10. Permanent Magnets and the Earth's Magnetic Field.—Permanent magnets are pieces of magnetically hard material which retain a considerable portion of the magnetic moment that they acquire on magnetization after the magnetizing field has been removed. The external magnetic effects due to a cylinder with a uniform axial magnetic moment are closely similar to those produced by a solenoid wrapped upon the curved surface of the cylinder. The amperian currents, which may be considered as responsible for the permanent magnetic effects, are also closely analogous to the persistent circulating currents associated with a material that has been rendered superconducting in a magnetic field. Electrets, which are permanently polarized dielectrics (Sec. 3.2), are a close electrostatic analogue of the permanent magnet from the pole point of view. The types of steels suitable for the construction of permanent magnets have been discussed in a previous section. Generally the shape in which the permanent magnet is made is determined by the purpose for which it is intended. However, when there is any latitude in the design, it should be so proportioned that the free poles subject the

bulk of the magnet to as little demagnetizing force as possible. The D'Arsonval galvanometer magnet is a particularly efficient design from this point of view.

The problem of using permanent-magnet materials to produce magnetic fields in air gaps can be solved in an approximately quantitative way by means of the concepts of the magnetic circuit [Eq. (11.13)]. Figure 11.18 illustrates schematically a section through a representative arrangement of blocks of permanent-magnet material so magnetized as to produce a permanent field across the air gap. The soft iron yoke or armature supports the blocks and supplies a flux path of small reluctance.

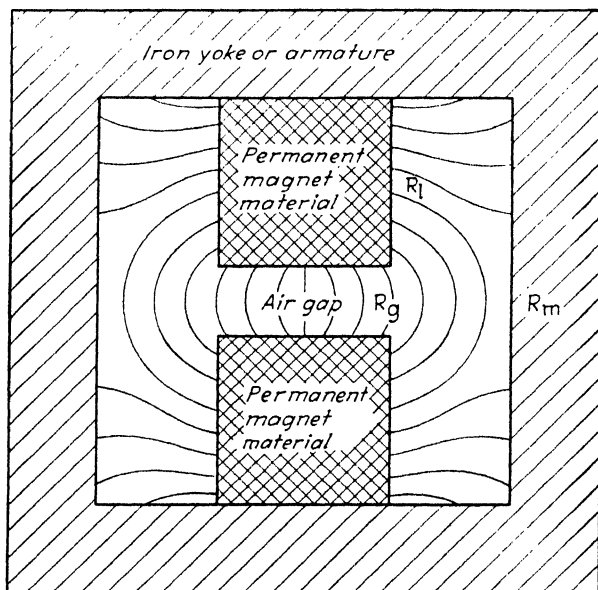



FIG. 11.18.—A permanent-magnet circuit.

If the subscripts p, m, g, and l refer, respectively, to the permanent magnet, iron, gap, and leakage, the total magnetomotive force supplied by p and the flux through p are related by $\mathcal{H}_p = \mathcal{R}\varphi_p$ where $\mathcal{R}$ is the reluctance of $\mathcal{R}_p$ and $\mathcal{R}_m$ in series with $\mathcal{R}_l$ and $\mathcal{R}_g$ in parallel. Also $\varphi_p = \varphi_l + \varphi_g$ and $\varphi_g\mathcal{R}_g = \varphi_l\mathcal{R}_l$. In consequence if $\mathcal{R}_m$ and $\mathcal{R}_p$ are small in comparison with $\mathcal{R}_g$ and $\mathcal{R}_l$, which is usually the case, $\mathcal{H}_g = \mathcal{H}_p$ and $\varphi_g = \mathcal{R}_l/(\mathcal{R}_l + \mathcal{R}_g)\varphi_p$. One is generally interested in a design that will minimize the amount of expensive magnet material for a given product of $\mathcal{H}_g\varphi_g$, which means a maximum H_pB_p, and a design that will reduce leakage flux and hence maximize $\mathcal{R}_l$. Assuming a negligible spreading of flux at the gap,

$$H_gB_g = \frac{\mathcal{R}_l}{\mathcal{R}_l + \mathcal{R}_g}\frac{l_pA_p}{l_gA_g}H_pB_p$$

where l and A represent length and area, respectively. Because of the factors mentioned in Sec. 11.9 it is difficult to increase $\mathcal{R}_l$ much above $\mathcal{R}_g$ by design; hence the factor involving the reluctance is at best of the order of $\frac{1}{2}$.

Figure 11.19 represents the upper left quadrant of an Alnico 5 sample and also the product of H_pB_p as a function of B_p. These are the curves of primary interest in permanent-magnet design. If the material is magnetized to saturation, on removal of the magnetizing coils it returns

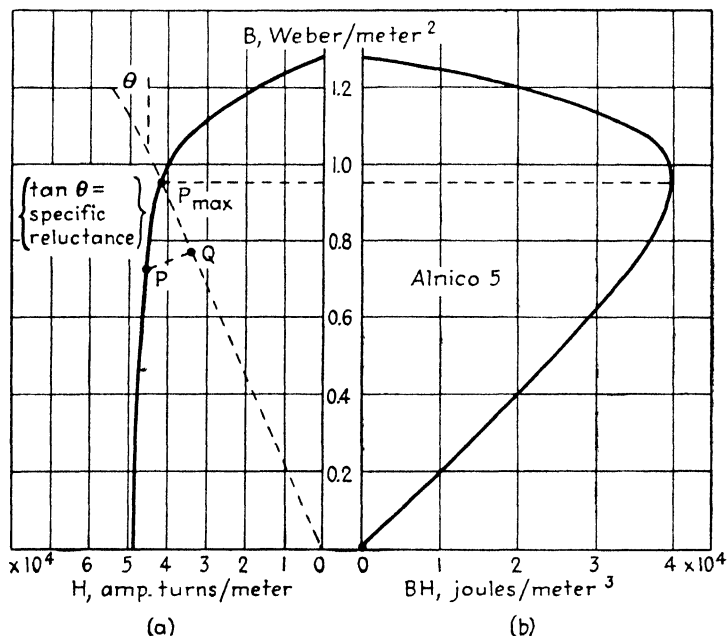



FIG. 11.19.—(a) Upper left quadrant of a saturation hysteresis curve showing the retentivity and coercive force for a permanent-magnet material. (b) The associated energy product (BH) curve.

to some point, say P_{max}, determined by the specific reluctance $\mathcal{R}A_p/l_p$ which depends on its dimensions and the reluctance of the magnetic circuit. In Fig. 11.19 P_{max} is the point corresponding to the maximum of the BH product which represents the most efficient utilization of the magnetic material. For stability it is well to apply a slight demagnetizing current beyond P_{max} so that the sample moves to the point P, and on removing this field the sample returns along a segment of a minor hysteresis loop (which in general is approximately parallel to the slope of the main hysteresis loop at $H = 0$) to the point Q. Subsequent small fluctuations in the effective H result in small motions of this point along the line PQ with but little change in the permanent equilibrium point Q as a result.

The uniformly magnetized sphere is an interesting calculable case illustrating a number of the phenomena associated with permanent magnets. If the magnetic moment per unit volume is constant throughout the sphere, the volume amperian currents vanish by Eq. (11.1) and the surface currents follow circles of latitude about the axis of magnetization. They are given in amplitude by

$$i_s^a = \frac{m_v}{\mu_0} \sin \theta$$

where θ is the polar angle. The magnetic scalar potentials to which this current distribution gives rise may be written

$$\Omega_i = Cr \cos \theta$$
$$\Omega_0 = \frac{D}{r^2} \cos \theta$$

inside and outside the sphere, respectively. The sphere is considered as replaced by this circulating current, the medium inside and out being free space. The conditions **curl H** $=$ $\mathbf{i}_v$ and div **H** $= 0$ imply that the normal components of **H** on the two

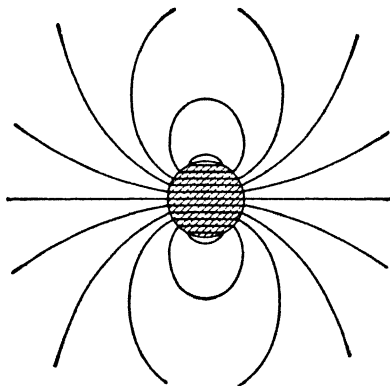



FIG. 11.20.—Lines of magnetic induction due to a uniformly magnetized sphere.

sides of the current sheet ($r = a$, where a is the radius of the sphere) are equal and that the difference between the tangential components is the surface-current density. These conditions determine the constants C and D and the potentials become

$$\Omega_i = \frac{2m_v}{3\mu_0} r \cos \theta$$
$$\Omega_0 = \frac{m_v a^3}{3\mu_0} \frac{1}{r^2} \cos \theta$$

The field inside the sphere, which is the negative gradient of Ω_i is thus $-\frac{2}{3}\frac{\mathbf{m}_v}{\mu_0}$ along the axis. The induction by Eq. (11.2) is $\mathbf{m}_v/3$. Thus the induction is in the direction of $\mathbf{m}_v$, but only one-third its magnitude, while the vector **H** is in the opposite direction. On writing **m** for the total moment, which is m_v times the volume, the external potential becomes

$$\Omega_0 = \frac{\mathbf{m} \cdot \mathbf{r}_1}{4\pi\mu_0 r^2}$$

This is the potential that would be produced by a current vortex of moment $\mathbf{m}$ at the center of the sphere and no external measurements can distinguish between a uniformly magnetized sphere and one containing a small magnet at its center.

The sphere is the only simply calculable case, but the general results obtained may be extended qualitatively to other shapes. The induction is always less than $\mathbf{m}_v$; it approaches this value for a very long axially magnetized cylinder and is less than $\frac{2}{3}\mathbf{m}_v$ for a cylinder short in comparison with its diameter. Empirical formulas giving the central induction for various shapes of cylinders and various types of magnetic materials are available in the literature.

A small motor operating on the Faraday disk principle can be made by mounting a magnetized rod so that it is free to rotate in a pair of vertical coaxial bearings. When a current flows from a brush contact on the curved surface, through the rod, and out at the bearings the magnetized rod is subjected to a torque. This is due to the force on the conduction electrons in the region of induction, say $\mathbf{B}$, inside the rod which is communicated to the magnet itself. The torque about the axis can be written

$$\mathbf{T} = \int \mathbf{r} \times (\mathbf{i}_v \times \mathbf{B})\, dv$$

where $\mathbf{r}$ is a vector perpendicular to the axis. The integrand is equal to $\mathbf{i}_v(\mathbf{r} \cdot \mathbf{B}) - \mathbf{B}(\mathbf{r} \cdot \mathbf{i}_v)$, and if $\mathbf{B}$ is axial, the first term vanishes. The second parenthesis is equal to the product of r and the radial component of $\mathbf{i}_v$ which for any coaxial radial shell is equal simply to the product of r and the total current i. Therefore

$$\mathbf{T} = -\mathbf{B}i\int_0^a r\, dr = -\frac{\phi i}{2\pi}$$

where ϕ is the total flux through the magnet. This is the same as the torque on a Faraday disk (Prob. 3, Chap. X).

The compass is simply a small permanent magnet that is pivoted in such a way that it is free to rotate in a horizontal plane. The equilibrium position of minimum potential energy is that in which the magnetic moment (South to North pole) is in the direction of the horizontal component of the field. Rigidly attached to the magnet is a pointer that moves over a scale graduated in degrees. This permits the measurement of the angular deflection of the magnet from any reference position. By comparison with astronomical observations the compass can be used to determine the angle between the horizontal component of the earth's field and the meridian. This angle is known as the *declination.* Magnetic maps or tables that give the declination for points of the earth's surface are of great importance for navigation. If a compass needle is mounted so that it is free to rotate in the plane of the magnetic meridian about a horizontal axis it will point in the true direction of the earth's field. A magnet so mounted is known as a dip needle and the angle

between the earth's field and the horizontal is known as the *inclination* or *angle of dip*. Local variations in the declination and dip are useful in locating deposits of ferromagnetic material.

The magnitude of a magnetic field such as that of the earth can be measured by making two different types of measurements with permanent magnets. First, a small magnet is suspended from its center by a fiber that exerts a negligible torsion so that it is free to oscillate in a horizontal plane. The torque exerted on it by the field, say $\mathbf{H}_e$, is equal to $\mathbf{m} \times \mathbf{H}_e$ or, since we are only concerned with its oscillation about a vertical axis, the restoring torque is $mH_e \sin\theta$, where θ is the angle between the moment and its equilibrium position and H_e is the

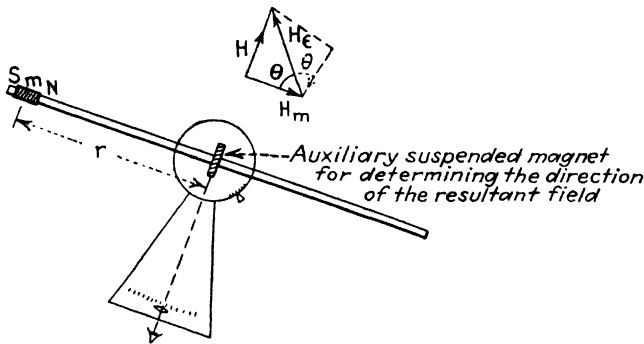



FIG. 11.21.—Magnetometer.

horizontal component of the field. If the magnet is set in small oscillation, its period, t_1, neglecting damping, is given by $2\pi\sqrt{I/mH_e}$, where I is its moment of inertia. If a mass of calculable moment of inertia I' is added, the period t_2 becomes $2\pi\sqrt{(I + I')/mH_e}$. Or, eliminating I between these expressions

$$mH_e = \frac{4\pi^2 I'}{t_2^2 - t_1^2} \tag{11.14}$$

Thus from a measurement of t_1 and t_2 the product mH_e is obtained. Now a small auxiliary test magnet is suspended by the fiber and the original magnet of moment m is placed on a frame that is capable of rotation about the axis of the fiber, in the horizontal plane of the suspended magnet and with its moment directed toward it. The arrangement is shown in Fig. 11.21. If r is the distance from the center of the moment m to the suspended magnet, the field at the latter point due to m is $m/2\pi\mu_0 r^3$ in the direction of m. When the test magnet is normal to m, it is seen from the construction that

$$H_m = H_e \sin\theta$$

or

$$\frac{m}{H_e} = 2\pi\mu_0 r^3 \sin\theta \tag{11.15}$$

where θ is the angle between the direction of m and the magnetic meridian. This setting of the auxiliary magnet is determined by viewing a scale in a small mirror attached to the magnet. The angle θ can be determined by reversing the direction of m but keeping it at the same point. As shown by the construction, a rotation of the apparatus through 2θ will result in the auxiliary magnet being again perpendicular to m. Measurements made at different values of r and with m on the opposite side of the test magnet eliminate errors due to instrument inaccuracies and the fact that m is not an ideal dipole. For these details the reader is referred to experimental treatises. The final result of these measurements yields a value for the ratio m/H_e. When both mH_e and m/H_e are known, m can be obtained by taking the square root of their product and H_e by taking the square root of their quotient. It should be pointed out that this is an absolute method of measuring either a magnetic moment or a magnetic field since only measurements of length, mass, and time are required.

All the factors that contribute to the earth's magnetic field are not well understood but the major aspects of the field are undoubtedly due to deposits of permanently magnetized ferromagnetic material in and beneath the earth's crust. Circulating currents in the atmosphere and the reception of charged particles from the sun also contribute a minor component. The surface field has long been an object of study and detailed maps giving its intensity and direction in practically all regions are available. Even the major features cannot be adequately represented by any simple function of the surface coordinates. However, to a very rough approximation the surface field is similar to one that would be produced by a magnetic dipole near the center of the earth or by a uniformly magnetized sphere coinciding approximately with the earth's surface. The magnetic axis is in the general direction of the polar axis which suggests a connection through the gyromagnetic effect between the earth's rotation and its magnetic moment. However, there is a considerable angle between these axes and also a displacement of the center of magnetism from the center of the earth. This is evidenced by the coordinates of the apparent terrestrial poles which are:

North pole: 70° 5′ N latitude; 96° 46′ W longitude
South pole: 72° 25′ S latitude; 155° 16′ E longitude

The maximum horizontal component of the earth's field is of the order of 0.4 gauss and is observed in the Philippine Islands. Through-

out the United States it is of the order of 0.2 gauss and the dip is in the neighborhood of 60 to 70°. The moment of the central dipole, which accounts approximately for the surface field, is about 10^{17} weber-meters. This may be expressed alternatively by saying that the earth has an apparent uniform magnetic moment per unit volume of about 10^{-4} weber per square meter. Superimposed on this permanent field is a comparatively small temporal variation. It is of a very complex nature, showing major periods of 1 day, 1 year, and 11 years. There are also minor periodic variations as well as secular ones and sudden changes

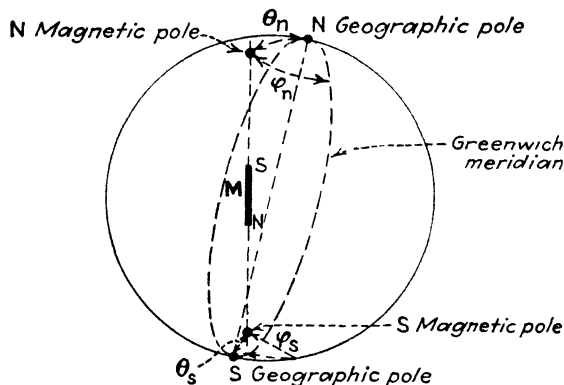



FIG. 11.22.—Location of the earth's magnetic poles and position of magnetic dipole that will approximately represent the field at the earth's surface.

known as magnetic storms which are probably associated with sun-spot activity.

Problems

1. A 600-turn toroid is wound on an iron ring of permeability 800, 20 cm. in diameter and 10 cm.² in cross section. Find the flux in the ring when the current is 1 amp.

2. Consider that the ring of the previous problem contains an air gap which is sufficiently short so that the induction in it can be considered equal to that in the iron. What is the value of the flux for air-gap lengths of 0.1, 1, and 5 mm.?

3. Calculate the field energy in the iron and in the gap, the total energy and the self-inductance for the ring of the previous problem in the case of each of the three gap widths.

4. Two long iron plungers of permeability μ are inserted in a very long closely fitting solenoid of n' turns per unit length. The cross section of the solenoid is A and it carries a current i. If the plungers are separated a short distance, show that the force tending to draw them together is

$$\frac{1}{2}\left(\frac{\mu - 1}{\mu}\right)\frac{AB^2}{\mu_0} \quad \text{or} \quad \frac{1}{2}A(\mu - 1)(n'i)^2\mu\mu_0$$

5. A U-shaped electromagnet has n turns of wire and carries a current i. A, l, and μ are its cross section, length, and permeability, respectively. If the pole separation is d, show that the force with which it retains a bar of the same cross section across its poles is approximately

$$F = \frac{A\mu_0\mu^2 n^2 i^2}{(l + d)^2}$$

Approximately how does this force vary with the gap if the bar is pulled a short distance away from the poles?

6. A circuit contains a toroidal coil 20 cm. in diameter and 5 cm.² in cross section which consists of 2,000 turns on a core of permeability 2,000. If the resistance of the circuit is 10 ohms, find the time constant, L/R.

7. Show that the mutual magnetic energy of two small magnets of moment $\mathbf{m}_1$ and $\mathbf{m}_2$ a distance $\mathbf{r}$ apart can be written

$$-\frac{1}{4\pi\mu_0}\mathbf{m}_2 \cdot \mathbf{grad}\left(\frac{\mathbf{m}_1 \cdot \mathbf{r}}{r_3}\right)$$

or

$$-\frac{m_1 m_2}{4\pi\mu_0 r^3}(\cos\phi \sin\theta_1 \sin\theta_2 - 2\cos\theta_1 \cos\theta_2)$$

where θ_1 and θ_2 are the angles between $\mathbf{m}_1$ and $\mathbf{m}_2$ and the vector from one to the other and ϕ is the angle between the two planes defined by the vectors $\mathbf{m}_1$ and $\mathbf{r}$ and $\mathbf{m}_2$ and $\mathbf{r}$.

8. Calculate the forces and torques on small coplanar magnets a great distance apart for the following cases: (*a*) colinear and parallel, (*b*) colinear and antiparallel, (*c*) parallel and perpendicular to r, (*d*) antiparallel and perpendicular to r, (*e*) one parallel and one perpendicular to r.

9. Show that if two small equal magnets are placed in a uniform field which is perpendicular to the line joining their fixed centers, they are in equilibrium when directed along the field if

$$H > \frac{3m}{4\pi\mu_0}\frac{1}{r^3}$$

where m is their magnetic moment and r their separation. (The condition for stability is that the magnetic energy U be a maximum, which implies that the coefficient of the second variation of the energy with respect to the coordinates is negative.)

10. Show that the energy of a small magnet of moment m in the presence of a long straight wire carrying a current i is $\mathbf{m} \cdot \mathbf{i} \times \mathbf{r}/2\pi r^2$, where $\mathbf{r}$ is the radial vector from the wire to $\mathbf{m}$. Find the force and torque on the magnet.

11. Two small magnets of moment m and mass M are suspended at their centers by light strings. If they are a great distance d apart, show that the magnets will attract one another and that in equilibrium the strings supporting them will be displaced from the vertical by the small angle

$$\theta = \frac{3}{2}\frac{m^2}{\pi\mu_0 d^4 M g}$$

where g is the acceleration of gravity. (The earth's field is neglected.)

12. A small magnet of moment m in a region of unit permeability is a distance a from a very large plane slab of material of permeability μ. Show that the magnetic energy is the same as that between two small magnets of moment m and m' a distance $2a$ apart for which $m' = \left(\frac{\mu - 1}{\mu + 1}\right)m$, and in the notation of Prob. 7, $\theta_1 = -\theta_2$, and $\phi = 0$ (method of images). Find the force of attraction between the magnet and slab at equilibrium if m is free to rotate.

13. A long straight wire carries a current i a distance a in front of a large plane slab of permeability μ. Show that the field in the slab is that which would be produced if the medium filled all space and a current $i'' = 2i/(\mu + 1)$ were in the place of the current i. Show that the field in the space in front of the slab is that which would be produced by i and another current, $i' = \dfrac{i(\mu - 1)}{(\mu + 1)}$ an equal distance beneath the surface of the slab. What is the force of attraction per unit length between the wire and the slab?

14. What is the difference in energy per unit length associated with a copper and an iron wire each of diameter 1 mm. and carrying a current of 1 amp. if the characteristic of the iron is linear and its permeability is 1,000?

15. Show that the lines of induction are effectively refracted at the boundary between two media in such a way that if α_1 and α_2 are the angles made by the lines with the normal to the boundary

$$\mu_1 \cot \alpha_1 = \mu_2 \cot \alpha_2$$

where μ_1 and μ_2 are the permeabilities of the two media.

16. A coil of wire is wound on an iron torus which exhibits a hysteresis curve such as that of Fig. 11.13. Assuming a sinusoidal current i, sketch the resultant flux as a function of the time. Assuming a sinusoidal flux, which corresponds to a sinusoidal emf. $\pi/2$ ahead of the flux, sketch the associated current.

17. Show from Eq. (10.3) that for a sinusoidal variation of flux the eddy-current losses per unit volume in a medium are proportional to the square of the frequency and inversely proportional to the resistivity. What can be said of the total losses as a function of the geometry?

18. Show that the magnetic scalar potential Ω [Eq. (11.4)] due to a long thin bar uniformly magnetized in the direction of its length reduces to

$$\Omega = -\frac{m_l}{4\pi\mu_0}\left(\frac{1}{r_1} - \frac{1}{r_2}\right)$$

at distances great in comparison with its cross-sectional dimensions. Here m_l is the moment per unit length, and r_1 and r_2 are the distances from the ends of the bar to the point at which the potential is Ω. (Compare with the potential due to two equal and opposite point charges.)

19. Show that to terms of the order of l^2/r^2 the potential due to such a magnet near its axis is

$$\Omega = \frac{-m}{4\pi\mu_0 r^2}\left[1 + \left(\frac{l}{2r}\right)^2\right]$$

Calculate the force between two colinear magnets to this approximation.

20. Show that to this approximation Eq. (11.15) for the magnetometer becomes

$$\frac{m}{H_e} = 2\pi\mu_0 \frac{r_1^5 \sin \theta_1 - r_2^5 \sin \theta_2}{r_1^2 - r_2^2}$$

where l has been eliminated by making two measurements of θ, namely, θ_1 and θ_2, for two positions of the magnet m at distances r_1 and r_2 from the auxiliary suspended magnet. (This correction is generally necessary since r/l is seldom greater than about 10 for deflections that can be measured with acceptable accuracy.)

21. Find the values of the magnetic induction both inside and outside of a long cylinder of a simple substance of radius a and permeability μ placed in a previously

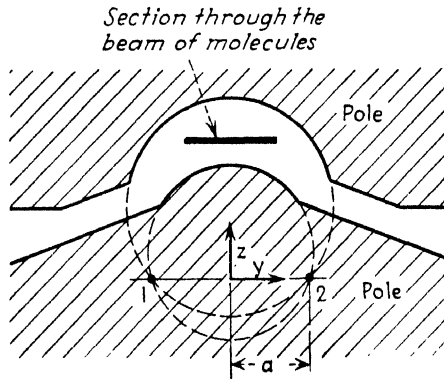
uniform field having a value $\mathbf{H}_0$ at a great distance. The axis of the cylinder is normal to $\mathbf{H}_0$.

22. Referring to Sec. 9.6 show that the self-inductance of a circular loop of wire of permeability μ is $\mu_0 R\left(\log \frac{8R}{r} - 2 + \frac{\mu}{4}\right)$, where R is the radius of the loop and r is the radius of the wire.

23. Show that the boundary conditions which obtain at a surface separating two magnetic materials 1 and 2 may be written in terms of the vector potential as

$$\mathbf{n} \cdot (\mathbf{A}_1 - \mathbf{A}_2) = 0, \qquad \mathbf{A}_1 - \mathbf{A}_2 = \mathbf{m}_s \times \mathbf{n}$$

where $\mathbf{n}$ is a unit vector normal to the bounding surface and $\mathbf{m}_s$ is the magnetic moment per unit surface area.

24. In the modification of the molecular beam apparatus used by Rabi the section through the pole pieces is not that shown in Fig. 10.16 but resembles that shown in

Prob. 24.

the accompanying figure. Assuming that the segments of cylinders forming the pole pieces are magnetic equipotentials, show that the y and z components of the field between them are given by

$$H_y = \frac{i}{2\pi} z\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right) \qquad \text{and} \qquad H_z = \frac{1}{2\pi}\left(\frac{d - y}{r_1^2} + \frac{d + y}{r_2^2}\right)$$

where r_1 and r_2 are the distances to the points 1 and 2, respectively. Determine the approximate expressions for the components of the gradients of the components of H on the assumption that y^2/z^2 is negligible in comparison with unity. Show that $\partial H_z/\partial z$ is approximately constant for $z \cong 1.2d$ and $0.7d > y > -0.7d$ and that the other components of the gradients are small in this region.

CHAPTER XII

ELECTROMAGNETIC MACHINERY

12.1. Introduction.—The most important practical application of electricity is the transportation of energy from one point to another. The primary source of the energy is the heat of the sun. By evaporation and precipitation surface water is given gravitational potential energy which may be converted into kinetic energy of rotation by means of hydraulic turbines. Also the biochemical processes induced by the sun's radiation produce fuel which may be used for running steam or mercury engines. These prime movers may in turn run electric *generators* which produce a large electromotive force and have a low internal resistance. They convert a large fraction of the mechanical energy into an electrical form in which it may be transported over wires to a distant point. There it may be reconverted into mechanical energy by a *motor*. Thus energy or power produced at one point may be distributed electrically and made available at a number of distant points. This is the most important economic function of electricity. Electrical engineering is the science dealing with the production, distribution, and reconversion of electric power. It is a vast and technical subject, but the fundamental principles upon which it is based have already been developed in the preceding chapters. The application of these principles to representative types of electrical machinery will here be considered briefly from the point of view of the student of electricity rather than that of the practicing electrical engineer.

In machinery of the type under consideration the electromotive force is produced by the relative motion of a conductor and a magnetic field. Continuous relative motion involves the use of rotating machinery and the natural introduction of a periodic time which is related to the period of rotation. Except in the case of the homopolar generator (in which the magnetic induction is parallel to the axis of rotation), the magnetic flux through the rotating member alternately increases and decreases cyclically, resulting in an electromotive force that alternates in the same manner. Thus an alternating current tends to flow in coils carried by the rotating member or *armature*, and if these are connected directly to the external circuit, an alternating current tends to flow in it. This type of machine is known as an *alternating-current generator*. However, the sense of connection of the armature coils and external circuit may be reversed at the proper time by means of an

automatic switch, known as a *commutator*, that is carried by the rotating member. In this way a unidirectional current of essentially constant magnitude may be produced. Such a machine is known as a *direct-current generator*.

Both alternating and direct currents are widely used for the transmission of power. Each system has its advantages and disadvantages. The difficulties associated with commutation are avoided in the alternating-current generator, but it is not self exciting or self-regulating. Direct-current motors also require a commutator and hence are in general of greater complexity than alternating-current motors, but the former have numerous advantages associated with greater flexibility in speed and torque control. It is in the distribution of power that the greatest difference exists between the two systems. In the first place, it is more efficient to transmit power at high voltages. The power transmitted, P, is Vi, and the joule loss in the lines is i^2R, where R is the transmission-line resistance. The ratio of the power lost to that delivered is thus $iR/V = PR/V^2$, which decreases as V increases. However, it is not convenient to reconvert electrical into mechanical power at high voltages. In the direct-current system the conversion requires the use of rotating machinery or electronic equipment whereas the transformer (Sec. 12.6) performs this function more efficiently and more economically in the alternating-current system. Thus the latter system has a distinct advantage from this point of view. However, the fact that the potential and current waves are not in general in phase leads to greater transmission losses than in the case of direct currents. The power delivered is $V_ei_e \cos \alpha$, where V_e and i_e are the effective values of the potential difference and current, respectively, and α is the phase angle (Sec. 5.5), whereas the line loss is i_e^2R, which does not involve the power factor ($\cos \alpha$). As the power factor is always less than unity, the fractional losses are greater than in the case of direct current. There are other technical difficulties associated with alternating-current distribution, but the ease of voltage transformation makes it distinctly superior for long-distance power transmission at present.

12.2. The Direct-current Generator.—The magnetic field in which the armature rotates is produced by an electromagnet, the winding of which is known as the *field*. Consider for simplicity that a region of uniform magnetic induction is produced in this way and that a frame of copper wire is rigidly attached to an axle perpendicular to **B**, as indicated in Fig. 12.1. The sides of length l are in the plane of the axle and parallel to it. The cross members of the frame which are of length d may be considered to lie outside the magnetic field as they do not play any part in the induction of the electromotive force. One of these terminal members is severed and the two ends are connected to the two

commutator segments that are insulated from one another and pass under the carbon brushes as the axle rotates. The most direct analysis of the action of the relative motion of the conductor and the magnetic induction is from the general force equation, Eq. (9.6). If no external electric field is present, the force on a conduction electron in the copper frame is $e\mathbf{u} \times \mathbf{B}$, where e is the electronic charge and $\mathbf{u}$ is the velocity of the frame carrying the electrons. The force per unit charge, or induced electric field, is thus

$$\mathbf{E} = \mathbf{u} \times \mathbf{B}$$

This field has no component parallel to the frame in the cross members even if they are in the field, but the field in the sides of length l is $\pm uB \sin \theta$,

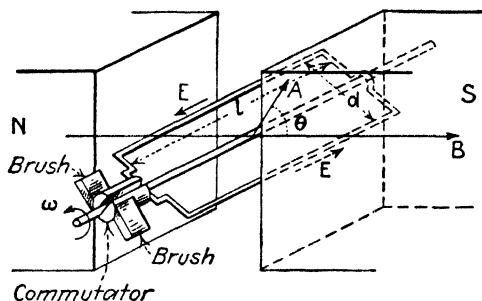



FIG. 12.1.—Conducting frame rotating in a magnetic field.

where θ is the angle between the normal to the plane of the frame and the induction $\mathbf{B}$. These fields are in opposite directions but in the same sense regarding the frame as a circuit, as indicated in the figure. Thus the integral of the electric field around the single-turn circuit, which is the electromotive force induced in it, is twice the product of E and l, or

$$\mathcal{E} = 2Blu \sin \theta \tag{12.1}$$

This may be written alternatively in terms of the angular velocity of rotation, $\omega = 2u/d$, or the frequency, $\nu = \omega/2\pi$, and the maximum flux, $\phi_m = Bdl$

$$\mathcal{E} = \omega\phi_m \sin \theta = 2\pi\nu\phi_m \sin \theta \tag{12.1'}$$

Since the circuit is a rigid one, except for the commutator which lies outside the field, it may also be considered from the point of view of Eq. (10.8). Integrating the **curl** of $\mathbf{E}$ over the surface of the frame is equivalent by Stokes's theorem to the line integral of $\mathbf{E}$ around the frame which is the induced electromotive force. As the integral of $\mathbf{B}$ over the surface of the frame is the normal flux through the coil, the induced emf. is the total rate of decrease of normal flux through the circuit. From the figure the normal flux through the circuit is $Bdl \cos \theta$, where

$\theta = \omega t$. Thus

$$\mathcal{E} = -\frac{d\phi}{dt} = \omega B d l \sin\theta = \omega\phi_m \sin\theta$$

in agreement with Eq. (12.1′).

The magnetic flux in which the armature rotates is produced by a current i_f flowing through the field coils which are wound on the pole pieces marked N and S in the figures. If there are m turns in the field, the magnetomotive force is mi_f and the flux is given by $\phi = mi_f/\mathcal{R}$, where $\mathcal{R}$ is the reluctance of the magnetic circuit. The frame of the machine, which is of iron or steel, provides a low-reluctance path for the flux, as shown in Fig. 12.7, except for the region in which the armature rotates. The reluctance of this gap is reduced by winding the armature on a steel drum with only a small clearance between it and the shaped

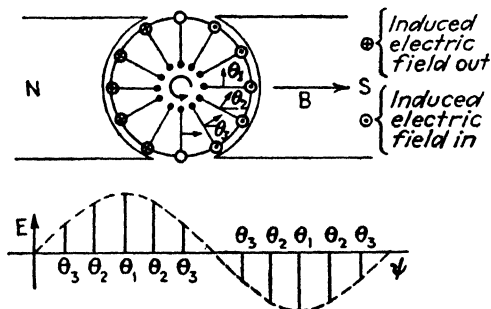



FIG. 12.2.—Induced electric field in the conductors of a drum armature.

pole pieces, as shown in Fig. 12.2. As this also rotates in the magnetic field, currents tend to flow in it parallel to the sides l of the armature. Hence this drum is built up of thin disk-shaped steel laminations which present a high resistance to axial current flow and minimize the eddy-current losses. The armature windings are recessed in slots on the surface of this drum.

It is evident from Eq. (12.1) that the potential difference between the brushes of Fig. 12.1 varies sinusoidally except for the periodic reversal by the commutator. Thus, although the emf. applied to the external circuit is unidirectional, it fluctuates between 0 and the maximum value $\omega\phi_m$. This fluctuation is reduced and the surface of the rotating drum is used more efficiently if a number of overlapping coils are disposed around the surface of the drum much as yarn is wound on a ball. A possible disposition of the windings in a six-coil armature are indicated in Fig. 12.6. The connections to the commutator which contains six segments are shown in Fig. 12.4. Since the connections between the armature and the external circuit are shifted by the commutator six

times per revolution instead of twice as in the case of the single frame. the angular motion of any one coil when connected to the external

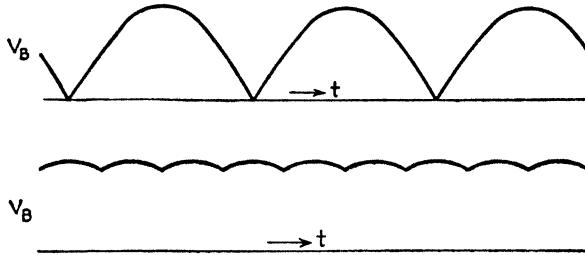



FIG. 12.3.—Brush potentials of simple direct-current generators. (a) Brush potential due to single coil of Fig. 12.1; (b) brush potential due to six-segment commutator of Fig. 12.4.

circuit is only $\pi/3$. If this occurs about the maximum of the sine function, the maximum fluctuation of the emf. is only 14 per cent. If the number of coils is increased, the fluctuation of the emf. is still further reduced.

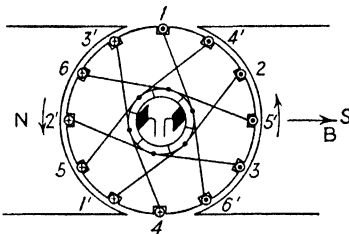



FIG. 12.4.—Drum-armature commutator connections.

Figure 12.2 represents schematically a section through such an armature. Assume for simplicity that the winding consists of a very great number of turns and that n_θ is the number connected in series between the brushes per unit angle around the armature. The emf. induced in the turns in the angular interval $d\theta$ is $n_\theta \omega \phi_m \sin \theta \, d\theta$ by Eq. (12.1′) and the total emf. is given by

$$\mathcal{E} = n_\theta \omega \phi_m \int_\pi^0 \sin \theta \, d\theta = n_\theta \omega \phi_m (\cos \theta)_{-1}^{1}$$

$$\mathcal{E} = 2 n_\theta \omega \phi_m$$

The total number of series turns, which will be written n, is $2\pi n_\theta$ and the emf. may be written in terms of the frequency of rotation as

$$\mathcal{E} = 2 n \nu \phi_m{}^* \qquad (12.2)$$

The circuits through the armature between the brushes are indicated more clearly in Fig. 12.5. The individual coils are here shown unwrapped from the drum and laid out in the order in which they occur.

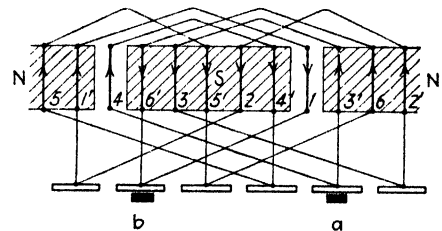



FIG. 12.5.—Developed diagram of the armature of Fig. 12.4 showing path under poles and instantaneous emfs. and brush positions.

* The average emf. in the case of a finite number of coils is also given by this expression.

The hatched path across them is that traversed by the magnetic flux. It is thus seen that there are two circuits in parallel through the armature. As the drum rotates, the coils in the figure occupy cyclically the successive positions in relation to the poles and brushes. The brushes are so disposed that the potential between them is a maximum, which implies that the line joining them is parallel to the induction **B**. From the nature of the sine function the rate of change of the potential between coils and hence between commutator segments is least where the potential difference is a maximum. Thus, though the brushes short-circuit adjacent commutator segments, the emf. thus short-circuited is a minimum and little sparking occurs for the correct adjustment.

Small generators are generally of the two-pole type indicated in the accompanying figures. Larger generators, however, have four or six poles and the armature windings may be more complex. For a detailed discussion of direct-current machinery the reader is referred to the technical literature.[1] The induced emf. is given by Eq. (12.2) and is constant if the speed of rotation and the flux remain constant. The frequency of rotation is generally limited by mechanical factors to some value between 10 and 60 r.p.s. and the induction is limited to about 1 weber per square meter by the properties of the iron or steel used in the magnetic circuit. The maximum potential developed is further limited by the commutator and insulation. The power put into the generator is, of course, the product of the torque T and the angular velocity ω, and the useful output is the product of the brush potential V and the current i. The iron, copper, and frictional losses account for the difference. The current that can be delivered is limited by a number of factors which depend principally on the heat developed within the machine and the electromagnetic reaction of the armature. Conservative practice is to limit the product of the current and number of conductors per centimeter of the armature periphery to about 100 for small machines, though this may be increased by a factor of 2 or 3 for very large generators.

FIG. 12.6.—Rear view of drum armature showing directions of armature currents and magnetic induction, A, due to them.

The effect of the current flowing in the armature is indicated in Figs. 12.6 and 12.7. From the sense of flow of this current it is evident that it gives rise to a magnetic field at right angles to the principal induction

[1] CREEDY, "Theory and Design of Electrical Machinery," Sir Isaac Pitman & Sons, Ltd., London, 1925; GRAY, "Electrical Machine Design," McGraw-Hill Book Company, Inc., New York, 1926; SLICHTER, "Principles of Design of Electrical Machinery," John Wiley & Sons, Inc., New York, 1926.

if the line joining the brushes is parallel to **B**. This induced induction is indicated by the arrow **A** in Fig. 12.6. The resultant induction is then in the direction **C**. The fact that the brushes are no longer in line with the induction causes a greater voltage across the commutator segments under the brushes, which results in sparking and incident losses. One method of correcting this is to rotate the brushes through a small angle in the direction of the armature rotation until the line joining them is parallel to the resultant induction. This arrangement is shown at the left in Fig. 12.7, but it is evident that the induced induction then has a component tending to reduce the total flux and the adjustment is strictly correct for only one value of the armature current. A still more important difficulty associated with commutation is that caused by the sudden reversal of the sense of current flow in the coils

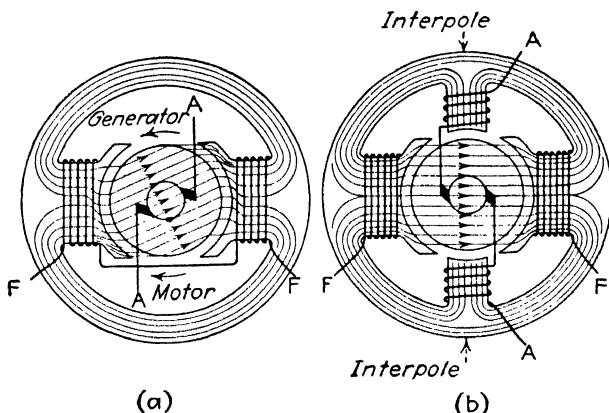



FIG. 12.7.—Armature reaction and its compensation. (*a*) Rotation of brushes to compensate armature reaction; (*b*) interpoles.

being commutated. Owing to the self-inductance of the circuit, an opposing emf. is induced which causes sparking at the brushes. This difficulty can be minimized by the introduction of one or more interpoles, as shown at the right in Fig. 12.7. These coils generally carry the armature current and are wound in such a sense as to induce an emf. in the coils being commutated that opposes the emf. induced by the changing current in these coils. The introduction of these poles obviously reduces the reluctance presented to the magnetomotive force of the armature and hence the windings on them must produce a considerably greater reverse flux than would be necessary in their absence.

Since the open-circuit brush voltage is proportional to the flux by Eq. (12.2), its dependence on the field current is given essentially by the reluctance of the magnetic circuit. At low flux densities the reluctance is largely that of the air gap, but as the flux density increases, the reluctance of the iron path becomes of more importance and the relation

between the brush voltage and field current resembles the magnetization curve. A typical curve giving the relation between these quantities is shown in Fig. 12.8. Owing to the residual or remanent magnetization of the iron circuit, the curve does not pass through the origin and owing to hysteresis there is a small difference between the curves, depending

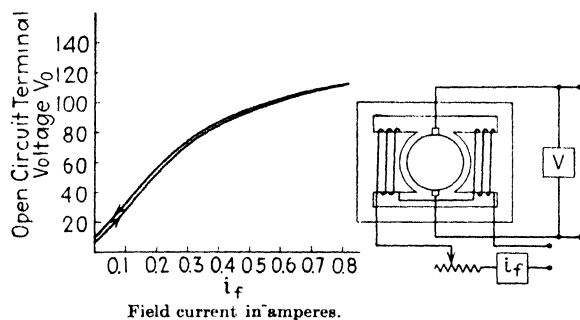



FIG. 12.8.—Typical magnetization curve of a direct-current generator.

on whether the current is increasing or decreasing. However, to a first approximation this difference can be neglected and the relation between V_0 and i_f may be taken as single-valued. Though the residual magnetization is small, it is of great importance if the current generated by the armature is used to excite the field. For in this case the generator begins to operate on this remanent induction and its direction determines the polarity of the machine.

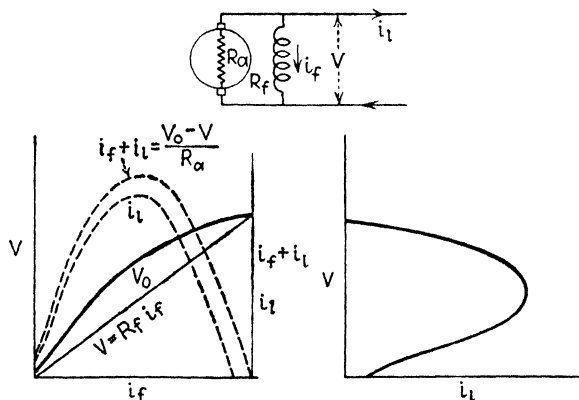



FIG. 12.9.—Shunt generator.

If the emf. generated by the armature is used to produce the field current, the machine is said to be *self-excited*. The field coils may be either in shunt across the brushes or in series with them and the load, or part of the field may be obtained in each way. In the latter case the machine is said to be *compound*. The voltage characteristic of the *shunt*

generator is illustrated in Fig. 12.9. The curves at the left indicate its graphical derivation. The curve marked V_0 is the open-circuit brush voltage as a function of the field current. Since the field current is proportional to the actual potential difference between the brushes, V, the latter is given by the straight line $V = R_f i_f$. The difference between these potentials V_0 and V causes the total current $i_l + i_f$ to flow through the armature of resistance R_a. Thus $i_f + i_l = (V_0 - V)/R_a$ and this total current is obtained from the difference of the two potentials at each point and indicated by the upper dashed curve of the figure. Finally the line current i_l is obtained by subtracting the abscissa from each point of the curve. From these curves the resulting values of V and i_l are plotted at the right. This is essentially the voltage characteristic of the machine. The only difference between this curve and

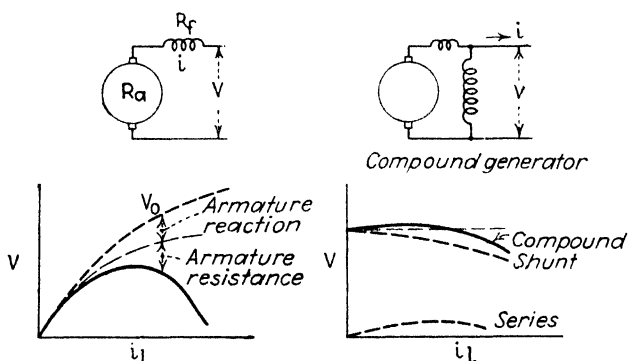



FIG. 12.10.—Regulation curves of series and compound generators.

those of Chap. V is that the potential is reckoned positive in the opposite sense, for this machine generates electrical power whereas the devices previously considered consume it. The lower portion of the curve having a negative slope is unstable and only the upper portion of the curve is traversed in practice. From this curve it is evident that the potential difference applied to the circuit terminals falls as the current rises, *i.e.*, as the load on the generator increases.

A series generator is indicated schematically at the left in Fig. 12.10. In this type of machine $V = V_0 - (R_a + R_f)i_l$, where the second term represents the iR drop in the generator. There is, of course, the drop in voltage due to the armature reaction in both types, and this is indicated explicitly in the series figure. It is evident that the series machine has a rising characteristic, *i.e.*, the terminal voltage increases with the load. The series generator is little used, but the rising characteristic associated with the series field is made use of to reduce the change in output voltage with load in the compound generator. In this machine part of the field is obtained, by a series coil, as indicated in Fig. 12.10. The shape of

the resultant characteristic depends on the proportion of the field obtained in the two ways. By a suitable choice the output voltage may be kept practically constant over a large fraction of the rated power output.

The Homopolar Generator.—If the disposition of the magnetic field and rotating member is such that the induction approaches the armature axially, a direct-current generator can be constructed without the use of a commutator. Such a machine is known as a homopolar generator. One type, which resembles the Faraday disk or Lorenz apparatus (Sec. 10.4), is indicated schematically in Fig. 12.11. An iron or steel cylinder is cut in half and a circular channel is milled in each of the opposing faces to contain the coils for producing the magnetic field. A disk is then mounted on a shaft that passes axially through the cylindrical block. The disk is made of steel to reduce the reluctance of the magnetic path and the shaft is mounted in bearings so that it may be rotated rapidly in the field. One electrical terminal is the shaft itself and the other is a sliding contact on the periphery of the disk. Owing to the high peripheral speed of the disk it is difficult to realize a satisfactory low-resistance contact to the external circuit. The return flux crosses the stationary lead from this contact and since there is no relative motion at this point, it contributes nothing to the emf. of the circuit.

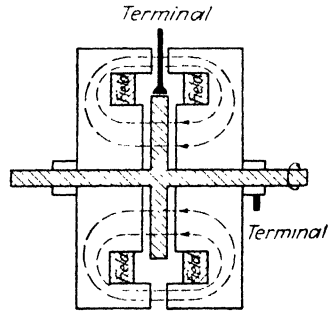



FIG. 12.11.—Faraday disk type of "homopolar" generator or motor.

Since the portion of the circuit on one side of this contact moves relative to the flux and that on the other side is stationary, the circuit cannot be analyzed as a rigid one from Eq. (10.8). However, the general force equation [Eq. (9.6)] can be applied directly, yielding an induced electric field at any point in the disk equal to $\mathbf{u} \times \mathbf{B}$, where $\mathbf{u}$ is the velocity of the disk at the point and $\mathbf{B}$ is the induction. If $\boldsymbol{\omega}$ is the vector angular velocity, $\mathbf{u} = \boldsymbol{\omega} \times \mathbf{r}$, where $\mathbf{r}$ is the radius vector from the axis. By the vector identity

$$\mathbf{E} = \mathbf{u} \times \mathbf{B} = -\mathbf{B} \times (\boldsymbol{\omega} \times \mathbf{r}) = \mathbf{r}(\boldsymbol{\omega} \cdot \mathbf{B}) - \boldsymbol{\omega}(\mathbf{r} \cdot \mathbf{B}) = \mathbf{r}(\boldsymbol{\omega} \cdot \mathbf{B}) = \mathbf{r}\omega B$$

since $\boldsymbol{\omega}$ and $\mathbf{B}$ are parallel. The total emf. between the axis and periphery is then

$$\mathcal{E} = \int_0^a \mathbf{E} \cdot d\mathbf{r} = \tfrac{1}{2}\omega B(r^2)_0^a = \tfrac{1}{2}\omega B a^2 = \pi a^2 B \nu = \nu \phi$$

where ν is the frequency of rotation and ϕ is the total normal flux through the disk of radius a. Since ϕ cannot easily exceed a fraction of a weber and ν is limited by mechanical considerations to about 50 r.p.s., the

emf. that can be generated in this way is only of the order of 10 to 30 volts. However, within this limitation such a generator performs very satisfactorily.

12.3. The Direct-current Motor.—A rotating machine for the reconversion of electrical power back into the mechanical form is called a motor. It is essentially the same type of machine as the electric generator but reversed in function. Any of the types of generators that have been discussed in the preceding section can be run backward and used as motors with only minor alterations. Electrical power is supplied to the terminals and mechanical power is obtained from the shaft. An instance of the type of alteration that may have to be made is that which corrects for the armature reaction. As the current flows through the armature of a motor in the opposite sense to that in a generator armature for the same direction of rotation (by Lenz's law), the armature reaction is in the opposite direction and the brushes must be shifted oppositely to the rotation to avoid sparking.

Consider the ordinary type of drum armature machine that was discussed at some length in the preceding section. If the armature is rotating at the frequency ν, the emf. generated is given by Eq. (12.2) as $2n\nu\phi_m$. If a current i is sent through the armature in opposition to this emf., the electrical forces do work at the rate $iV = 2ni\nu\phi_m$ which is converted into mechanical power and maintains the rotation against an opposing torque. If this torque is T, the rate of generation of mechanical power is $T\omega = 2\pi\nu T$. Equating the power in electrical and mechanical terms and writing the torque explicitly

$$T = \frac{ni\phi_m}{\pi} \tag{12.3}$$

As an exercise the same equation can be derived directly from the force equation applied to the drum armature carrying a current i in the region of constant induction B. Of course, the electrical reaction (*i.e.*, the back electromotive force) is not quite equal to that applied to the motor, for there are some losses which may be considered chiefly as due to the iR drop in the windings of the machine. Thus, if V is the potential applied to the motor terminals

$$V - V_b = iR$$

where V_b is the back emf. and R is the effective ohmic resistance of the machine. From Eq. (12.2) the back emf. is $2n\nu\phi_m$ and the equation may be written

$$\nu = \frac{V - iR}{2n\phi_m} \tag{12.4}$$

The second term in the numerator is essentially a small correction term for an efficient motor. The fraction of the power delivered to the motor that is converted into mechanical power is the efficiency. For small motors it is of the order of 75 per cent and for large ones it is as high as 90 per cent. The ratio of iR to V, which is the fraction of the power lost (if i^2R represents the total losses), thus lies in the range from 0.1 to 0.25.

Equations (12.3) and (12.4) are the basis for the simple discussion of motor performance. The torque is directly proportional to the current and the flux. The speed, on the other hand, is a linear function of the current but decreases as the current increases and is inversely proportional to the flux. Assuming the line voltage to be constant, the flux is also constant and hence the torque increases linearly with the current and the speed decreases linearly. This decrease, however,

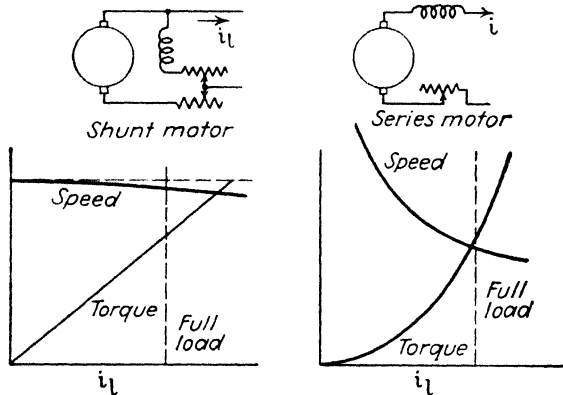



FIG. 12.12.—Direct-current motor characteristics.

is small and the total power output increases approximately in proportion to the current. The *shunt motor* is essentially separately excited for the field coils are directly across the line and hence the current in them is to a first approximation independent of that in the armature. The variation in speed and torque with line current is shown at the left in Fig. 12.12. It is evident that the speed varies little with the load, which is proportional to the product νT from no load to well beyond the rated output of the machine. Thus this type of motor is particularly suited to constant-speed operation. The sense of rotation is unaltered by reversing the external terminals but can be changed by reversing either the armature or field terminals separately. The speed may be increased by reducing ϕ_m, which may be accomplished by increasing the resistance of a rheostat in series with the field. Increasing the armature current decreases the speed somewhat but increases the torque. The double-rheostat type of control indicated schematically in Fig. 12.12, which

changes the armature and field currents in opposite senses, is seen from the equations to yield approximately the same speed control with a smaller change in torque. For starting a motor of this type a series resistance should be included in the armature circuit. Owing to the mechanical inertia of the rotating parts, the back emf. is not established at once, and while the motor is getting up to speed, an excessive current would be drawn by the armature windings. The gradual increase of current obtained by reducing the resistance in series with the armature protects the load from mechanical shock and the distributing system and protective devices from electrical surges.

The *series motor*, which is indicated at the right in Fig. 12.12, possesses entirely different characteristics. In this type the field and armature windings are in series as the name implies and hence i_f and i_a are equal to one another and to the line current. In the ideal case in which ϕ_m is the total magnetic flux produced by the field and the magnetization of the armature current is neglected, $\phi_m = mi/\mathcal{R}$, where m is the number of turns in the field and $\mathcal{R}$ is the reluctance of the magnetic circuit. Substituting this value of ϕ_m in Eqs. (12.3) and (12.4) and neglecting the iR term the speed and torque are given by the following equations:

$$\nu = \frac{V\mathcal{R}}{2nmi}. \qquad \text{and} \qquad T = \frac{nm}{\pi\mathcal{R}}i^2$$

The torque increases parabolically with i instead of linearly as in the case of the shunt motor, and assuming V to change but little ν is inversely proportional to i instead of practically constant as in the case of the shunt motor. Eliminating i between these parametric equations, T is seen to be proportional to V^2 and inversely proportional to ν^2. Thus, when ν is small as in starting, a very large torque is developed, but if the load is removed so that the torque is small, ν tends to become very great and the excessive speed developed may damage the machine unless it is limited in some way as by a small shunt field. The high starting torque and variable-speed characteristics are particularly suited to traction work. Reversal can be accomplished by reversing either the armature or field separately but not by reversing the line terminals as this leaves the sense of the torque unchanged. At constant torque the speed is seen to be proportional to the potential V applied to the armature which to a first approximation is equal to that applied to the motor terminals. Speed control can be obtained by means of a series rheostat which also affords starting protection.

12.4. The Alternating-current Generator.—The ordinary type of alternating-current generator, or alternator, resembles the direct-current generator in the principles of its construction. A magnetic flux is produced by a direct current flowing in field coils and the magnetic

circuit is as nearly closed as possible to reduce the reluctance of the path. However, the details of the armature windings and the connections to them are quite different. In the simpler type of machine there is only one armature coil per pole pair per phase and the two terminals associated with the phase are connected to two slip rings that are coaxial with the shaft if the armature is the rotating member. The generated emf. is applied to the external circuit through two stationary brushes in contact with the rings. This type of machine is represented schematically in Fig. 12.13. If n is the number of turns in the coil shown

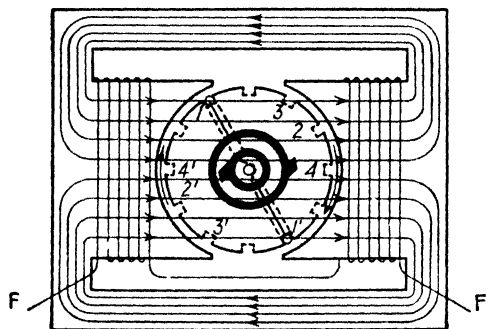



FIG. 12.13.—Schematic alternating-current generator or synchronous motor.

in the slots 1 and 1', the emf. generated is given ideally by Eq. (12.1) or Eq. (12.1') as

$$\mathcal{E} = 2nBlu \sin \theta$$

or

$$\mathcal{E} = \pi n\nu\phi_m \sin 2\pi\nu t \tag{12.5}$$

Thus the potential difference between the brushes when no current is drawn is given by this expression which is plotted as a function of the time in the upper portion of Fig. 12.14.

If another coil had been wound on this armature, say in the slots 2 and 2', and the terminals brought out to another pair of slip rings, the potential difference between these rings would be given by a similar expression, except the characteristic features of the curve would occur at a later time equal to one quarter of the period of revolution. This means that there would be a phase lag between the two sine curves of $\pi/2$. The potential difference applied to the second set of brushes would then be

$$V_2 = 2\pi n\nu\phi_m \sin\left(2\pi\nu t + \frac{\pi}{2}\right) \tag{12.6}$$

This curve is indicated by the dashed wave about the central axis of Fig. 12.14. Such a machine supplying two sine waves is known as a *two-*

phase or *quarter-phase* generator. If this winding were omitted but the slots 3 and 3′ and 4 and 4′ were wound and these coil terminals brought out to two additional slip rings, the three sine waves thus generated would be

$$V_1 = V_0 \sin 2\pi\nu t$$
$$V_2 = V_0 \sin\left(2\pi\nu t + \frac{2\pi}{3}\right)$$
$$V_3 = V_0 \sin\left(2\pi\nu t + \frac{4\pi}{3}\right) \qquad (12.7)$$

where V_0 is written for $2\pi n\nu\phi_m$. These are shown in the lower portion of Fig. 12.14, and the machine with these three windings is known as a

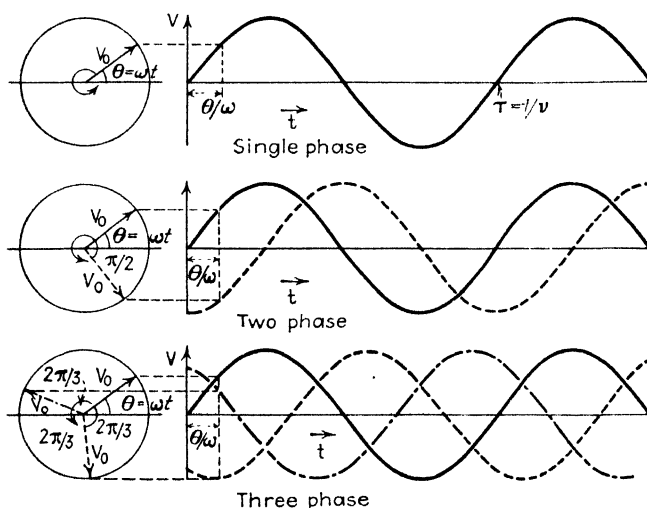



FIG. 12.14.—One-, two-, and three-phase alternating-current potential waves.

three-phase generator. All three types are in use though large modern machines are generally of the three-phase type. The rotating-vector representations of these potential waves are shown at the left in Fig. 12.14. If V_0 is the length of a vector rotating with an angular velocity ω, the vector **V** representing its magnitude and direction at any instant would be given by $V_0e^{j\omega t}$. The projection of this vector on the vertical or imaginary axis is then the imaginary part of **V** which is $V_0 \sin \omega t$ or $V_0 \sin 2\pi\nu t$. The horizontal lines connecting the pairs of diagrams indicate that the sine waves drawn are the projections of the vectors on the vertical axis.

In large machines the relative positions of the field and armature are generally reversed. The constant field rotates within stationary armature coils carried by the frame of the alternator. In this arrange-

ment the stationary iron magnetic circuit upon which the armature coils are wound must, of course, be laminated as they are subjected to a changing flux which induces eddy currents. The production of a constant field requires a direct current for the windings. This is usually supplied by a separate generator. As the rotating machine does not supply its own field, it lacks the self-regulating feature of the direct-current generator, but electrical interconnection and automatic voltage regulation may be employed to keep the output voltage maxima approximately constant independent of the load. The field current must, of course,

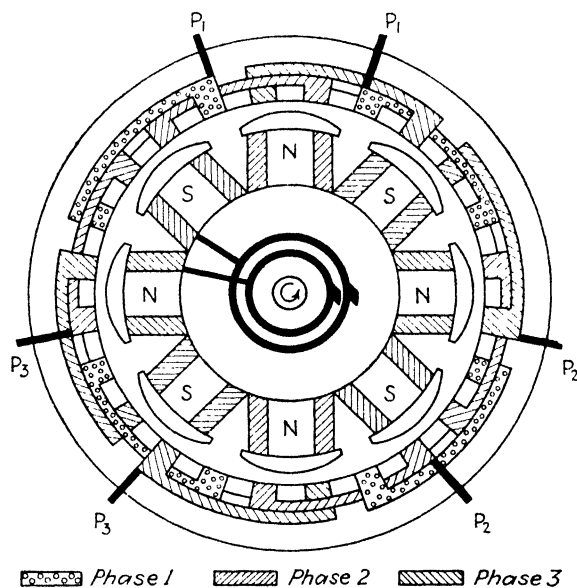



FIG. 12.15.—Schematic three-phase alternator or motor.

be brought into the rotating member by means of slip rings. The power delivered to the generator is the product of the torque and the angular velocity and that converted into a useful electrical form is the product of the generated potential and the line current. Of course, the power fluctuates throughout the cycle, having an average value of $V_e i_e \cos \alpha$, where subscript e indicates the effective value (the maximum value divided by $\sqrt{2}$) and $\cos \alpha$ is the power factor which is largely determined by the load (see Sec. 5.3). The effective value of the potential is given by the preceding equations as $\sqrt{2}\pi\nu\phi_m$ or $4.44\nu\phi_m$. The permissible power output of the machine is determined principally by the product of the current per conductor and the number of conductors per centimeter of the armature periphery, *i.e.*, by the current density on the armature periphery. This may lie between about 100 for small machines and a maximum of about 500 for large ones.

The electrical frequency is, of course, equal to the product of the frequency of rotation and the number of pole pairs. For the eight-pole machine represented in Fig. 12.15 the electrical frequency is 4ν. That is, the machine would have to rotate 15 times per second or 900 times per minute to generate the ordinary commercial 60-cycle frequency. There are four coils in series for each phase, one for each pole pair. The three separate circuits are distributed equidistantly around the armature periphery as indicated in the figure. One quarter of the total emf. is generated in each of the four coils of the series set. The wave shape depends, of course, on the flux distribution over the air gap between field and armature. This is generally controlled by the shape of the pole face. An approximately sinusoidal wave is obtained if the pole face is an arc of a circle such that the gap is approximately twice as great at the pole tips as at the center. When a current is drawn from the armature, there is of course an armature reaction as in the case of the direct-current machine, though its effects can be minimized by proper design. For a further discussion of alternator design reference should be made to the technical literature previously cited.

12.5. The Synchronous Motor.—As in the case of the direct-current machines, an alternator can be reversed in function and run as a motor. Such a machine is known as a synchronous motor, for it rotates at such a speed that its electrical output, considering it as a generator, would be of the same frequency as the potential wave actually applied to its terminals. As in the case of the alternator, it requires direct-current excitation to produce the constant magnetic field. A motor of this type has several electrical advantages. As its speed is determined by the electrical frequency (being equal to the frequency divided by the number of pole pairs) it is constant independent of the line voltage or load. Thus it may be used to run a direct-current generator for obtaining a constant output voltage, or any other device that must be run at a constant speed. Also the phase angle between the current through the machine and the voltage applied to its terminals can be varied by varying the direct exciting current that produces the magnetic field. For a large value of this exciting current the alternating-current wave is in advance of the voltage wave. Thus it acts as a condenser (Sec. 7.6) and it may be used to compensate for the inductive reactance that is presented by another type of load such as an induction motor. If the two machines are run from the same line, the power factor can be made unity with a consequent higher efficiency in the distributing lines.

If an effective current i_e flows to the motor against an effective potential V_e, the electric power consumed is $i_eV_e \cos \alpha$. The mechanical power produced is $T\omega$ and, neglecting losses, these are equal. The synchronous potential produced by the motor is given by Eq. (12.5);

hence the torque developed is

$$T = \frac{n\phi_m}{2\sqrt{2}} i_e \cos \alpha \tag{12.8}$$

This may be considered vectorially by the aid of Fig. 12.16. $\mathbf{V}_g$ is the vector representing the generator voltage that is applied to the motor. $\mathbf{V}_m$, which is shown at the angle $\pi - \beta$, is the back emf. generated by the revolution of the synchronous motor. It is shown in the diagram somewhat shorter than V_g, though this depends on the excitation. The excitation corresponding to the vectors as shown is that which would make the output voltage of the motor considered as a generator less than the generator actually present in the line. The resultant voltage $\mathbf{V}$ sends the current $\mathbf{i}$ through the armature. When there is no load, $\mathbf{V}_g$ and $\mathbf{V}_m$ are antiparallel and the net voltage $\mathbf{V}$ is very small. As the load is increased, the electrical phase difference $(\pi - \beta)$ between the motor and generator changes so that the resultant $\mathbf{V}$ grows and the requisite power is supplied. The locus of $\mathbf{V}$ with varying load is evidently a circle of radius V_m about the extremity of $\mathbf{V}_g$ as a center, as shown in the diagram. The impedance presented by the armature is almost entirely reactive so that the phase of the vector $\mathbf{i}$ is practically $\pi/2$ behind $\mathbf{V}$ as shown. The length of $\mathbf{i}$ is the quotient of V_0 and the generator impedance. Since the latter is practically constant, the locus of $\mathbf{i}$ with varying load is also a circle as shown. The mechanical power, neglecting losses, is $V_{eg} i_e \cos \alpha$, where α is the angle between $\mathbf{i}$ and $\mathbf{V}_g$. It is a maximum when $\mathbf{i}$ has the greatest component along the $\mathbf{V}_g$ axis, as shown in the figure. If the load increases beyond this point, the angle α decreases further, but the power input evidently decreases and the motor slows down and stops.

FIG. 12.16.—Vectors for a synchronous motor and generator (maximum power output shown).

FIG. 12.17.—Locus of synchronous motor vectors as a function of the exciting field current (constant load).

The effect of varying excitation at constant load is shown in Fig. 12.17. The vector $\mathbf{V}_g$ representing the alternating potential of the generator is constant and the projection of the current vector $\mathbf{i}$ on $\mathbf{V}_g$ is also constant since the speed is invariant and the torque is assumed to remain the same. Thus the locus of the tip of the current vector is a straight line perpendicular to $\mathbf{V}_g$. $\mathbf{i}_1$ represents a current vector lagging by

the phase angle α_1. $\mathbf{V}_1$, which is the net emf. in the circuit, is approximately perpendicular to the current vector and proportional to it in length. Thus the locus of the tip of this vector is a straight line parallel to $\mathbf{V}_g$. The vector difference between $\mathbf{V}$ and $\mathbf{V}_g$ is the back emf. of the motor $\mathbf{V}_m$. If the group of vectors is considered to rotate about the origin, $\mathbf{V}_m$ should of course be drawn from that point, but in the diagram it is displaced so as to complete the triangle in order to bring out the relation between the three potential vectors. If $\mathbf{V}_m$ is less in magnitude than $(V_{0g}^2 + V_0^2)^{1/2}$, it is evident that the current lags behind the generator voltage, as indicated by the vectors with the subscript 1. If, however, $\mathbf{V}_m$ is greater in magnitude than this, the current leads the generator voltage as indicated by the vectors with the subscript 2. Since the speed is constant, the back emf. of the motor is determined by the field strength and hence an increase in the field current causes the alternating-current vector to advance in phase.

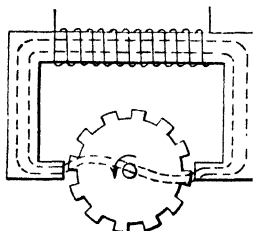



FIG. 12.18.—Small hysteresis type of synchronous motor.

The strict dependence of the speed of a synchronous motor on the electrical frequency makes it a particularly valuable device for the mechanical registration or counting of regularly spaced electrical impulses. The angular rotation of a synchronous motor is directly proportional to the total number of current waves that have passed through it. Thus, if a constant frequency is supplied to the motor, it can be geared to actuate an ordinary clock mechanism and register time. The accuracy of the clock depends, of course, on the accuracy with which the constancy of the electrical frequency is maintained. If the electrical impulses are generated by the mechanical oscillations of a carefully controlled quartz crystal (Sec. 14.3) and a submultiple frequency is selected by means of a multivibrator circuit (Sec. 15.7) to run the synchronous recorder, the device constitutes a more accurate clock than can be constructed mechanically. The ordinary alternating-current power lines may also be used to run clocks if the generator frequency is carefully controlled. Clock motors require so little power that it is not necessary to supply separate field excitation. A permanent magnet may be used as the rotating-field member. Alternatively the lag due to hysteresis in the field member or a variation in the reluctance of the flux path may be used to produce the torque. Figure 12.18 represents schematically a simple motor of the clock type. The alternating current flows in the coil wound on the yoke and the wheel tends to take up a position in which a pair of protuberances are in line with the armature poles. As the armature polarity changes, that of the wheel tends to

remain the same owing to the magnetic characteristics of the iron, and a repulsion results between the armature poles and the remanent poles of the wheel. By the time the next pair of protuberances have come under the armature poles the mmf. has risen so that these in turn assume the opposite polarity of the armature and are repelled during the ensuing half cycle. Thus once the motor is started, it will run in synchronism with the applied potential wave. In common with all other synchronous motors it is not self-starting but must be given an initial impulse or some other motor principle employed to bring it up to synchronous speed. The sense of rotation of a single-phase synchronous motor depends on the direction in which it is started; it will run equally well in either sense. Two- and three-phase motors, however, are unidirectional owing to the sequence of the windings, as will be brought out more clearly in the discussion of the rotating magnetic field.

12.6. The Transformer.—Though the transformer is not an electro-mechanical machine, it is such an important element in alternating-current power circuits that its simple theory will be considered at this point. Its general theory will be considered in more detail in connection with coupled circuits (Sec. 14.1). The most important service performed by the transformer is to change the potential-current ratio at which the electric power is delivered, *i.e.*, the product i_eV_e is unchanged (neglecting losses) but the ratio i_e/V_e is altered by the transformer. A typical transformer consists of a closed iron path about which are wrapped two windings of wire called the *primary* and *secondary*. Two schematic dispositions of the iron core and windings are shown in Fig. 12.19. The primary is connected to the source of alternating-current power and the power output is derived from the secondary terminals. The two types shown in the figure perform essentially the same function. The core type is better suited to high-voltage circuits as the insulation can be accomplished more simply; but there is in general less leakage flux in the case of the shell type. That is to say that there is less flux that does not follow the iron path and link both coils, *i.e.*, the square of the mutual inductance between them is more nearly equal to the product of the self-inductances (Sec. 9.5). In the actual construction of power transformers of the core type the primary and secondary windings are contiguous and each is wound on both legs of the core. Transformers resembling the schematic diagram are, however, occasionally used when a large leakage flux is desired. In the ideal transformer that will be

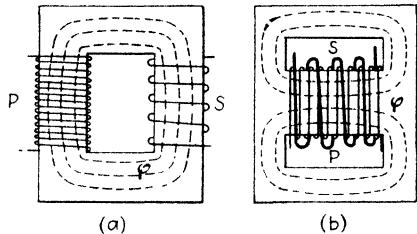



FIG. 12.19.—Schematic representations of transformers. (*a*) Core type; (*b*) shell type.

here considered the leakage flux is negligible and the ohmic resistance of the windings will also be neglected.

Consider first that the transformer is a strictly ideal linear device and neglect for the moment the presence of the secondary winding. If a current i_p flows in the primary, it will produce a flux given by

$$\phi = \frac{\mathcal{H}}{\mathcal{R}} = \frac{n_p i_p \mu \mu_0 A}{l} \tag{12.9}$$

where n_p is the number of primary turns, μ is the permeability of the iron path, A is the core area, and l is its effective length. If i_p is a sinusoidal function of the time, ϕ will be also. This changing flux, of course, implies that eddy currents are induced in the iron core, and this must be laminated to prevent excessive losses due to these circulating currents. The back emf. induced in the primary winding is $-n_p \frac{d\phi}{dt}$ and as i_p and ϕ are assumed to be sinusoidal functions of the time ($\mathbf{i}_p = i_{p0}e^{j\omega t}$ and $\phi = \phi_0 e^{j\omega t}$)

$$\mathbf{V}_{pb} = -j\omega n_p \phi = -j\omega\left(\frac{n_p^2 \mu\mu_0 A}{l}\right)\mathbf{i}_p \tag{12.10}$$

This must be equal and opposite to the potential applied to the primary if there are no losses. The quantity in brackets is the self-inductance of the primary circuit. If now there is a secondary winding the emf. induced in it is evidently equal by the same argument to $-j\omega n_s \phi$, where n_s is the number of turns in the secondary circuit. From this it is seen that the ratio of the secondary to primary emfs. is

$$\frac{\mathbf{V}_s}{\mathbf{V}_{pb}} = \frac{n_s}{n_p} \tag{12.11}$$

This is said to be the *ratio of transformation*. Since by hypothesis the losses are negligible, $\mathbf{V}_p\mathbf{i}_p = \mathbf{V}_s\mathbf{i}_s$, and the currents are given by

$$\frac{\mathbf{i}_s}{\mathbf{i}_p} = -\frac{n_p}{n_s} \tag{12.12}$$

since $\mathbf{V}_{pb} = -\mathbf{V}_p$. The ratio $\mathbf{V}_p/\mathbf{i}_p$ is the effective impedance presented by the primary which is given by

$$\frac{\mathbf{V}_p}{\mathbf{i}_p} = \left(\frac{n_p}{n_s}\right)^2 \mathbf{z}_s \tag{12.13}$$

where $\mathbf{z}_s$ is the impedance of the secondary circuit, $\mathbf{V}_s/\mathbf{i}_s$. Thus the impedance presented by the load at the primary terminals of the transformer is equal to the actual impedance divided by the square of the ratio of transformation.

The preceding equations describe the principle features of the ordinary power transformer. The potential is increased by the transformer in the ratio of transformation and the current is reduced in this ratio. By hypothesis there is no change in the product, *i.e.*, no loss of power. $\mathbf{V}_p$ and $\mathbf{V}_s$ differ in phase by π as do $\mathbf{i}_p$ and $\mathbf{i}_s$; furthermore by Eq. (12.13) the power factor of the primary is the same as that of the secondary. As the flux linking the primary and secondary circuits is the same, the sign of the induced emf. in the secondary is the same as that in the primary if the windings are taken in the same sense. This

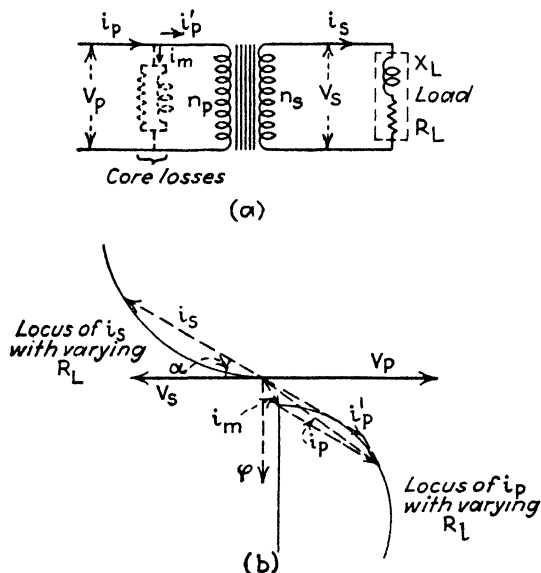



FIG. 12.20.—(*a*) Circuit of a transformer neglecting resistance and leakage inductance of the windings. (*b*) Vector relations in an ideal transformer.

choice of sign is particularly obvious if there is a conducting connection between the two circuits. Thus, if one coil is a continuation of the other and one terminal of each winding is common, the emf. developed in each would be reckoned positive in the same sense and the vectors would be parallel. A transformer of this type that consists essentially of a single winding tapped, say, at a fraction f of the turns is known as an *autotransformer*. If the fraction f of the turns constitutes the primary and the secondary is taken from across the entire winding, the transformation ratio is $1/f$. Such a transformer is frequently used to effect a small increase in voltage to compensate for a line drop.

A well-designed transformer has a very high efficiency ($>$ 95 per cent), but there are two small sources of loss, one associated with the core and the other with the windings. The core losses are due to hysteresis and eddy currents and may be minimized by choosing a steel having

a high resistivity and a hysteresis loop of small area. The losses increase with the thickness of the laminations and with the flux density. The flux lags the exciting current that produces it and the effect of the core can be represented by a hypothetical resistance and reactance in shunt as shown in Fig. 12.20. The vector diagram of the same figure indicates the so-called magnetizing current $\mathbf{i}_m$ that flows in this circuit. $\mathbf{V}_p$ represents the potential *applied* to the primary; the back emf. induced in the primary and the emf. induced in the secondary are in the opposite direction. If the secondary circuit is open, $\mathbf{i}'_p$ and $\mathbf{i}_s$ are zero and the difference between the applied potential and back emf. is just sufficient to produce the magnetizing current. If the secondary circuit is closed, a current $\mathbf{i}_s$ flows in it that is determined in magnitude and phase angle by the load. An inductive load resulting in a phase lag α is shown in the figure. If the load resistance is varied, the locus of the tip of the secondary current vector is a circle of diameter V_s/X_L tangent to V_s at the origin. This is evident from Sec. 10.1, for

$$\dot{\mathbf{i}}_s = \frac{\mathbf{V}_s}{\mathbf{z}_L}\quad\text{or}\quad \dot{i}_{0s} = \frac{V_{0s}}{Z_L} = \frac{V_{0s}}{X_L}\frac{X_L}{Z_L} = \frac{V_{0s}}{X_L}\sin\alpha$$

where α is the phase angle and V_{0s} is the magnitude of the secondary potential. This is the equation of the circle referred to.[1] The primary winding current $\mathbf{i}'_p$ is proportional to $\mathbf{i}_s$ but in opposition to it. The sum of $\mathbf{i}'_p$ and $\mathbf{i}_m$ is the primary line current $\mathbf{i}_p$, the locus of which with changing secondary resistance is evidently also the circle shown. Thus the phase lag of the primary current is greater than that of the secondary by a small amount due to the magnetizing current. The complete theory of the transformer would, of course, include the effects of the resistance and leakage inductance of the windings. These would result in small additional corrections, but they will be considered from a different point of view in Sec. 14.1.

In order to keep the core losses small, as assumed in the preceding discussion, the maximum value of the flux density (B) must not be allowed to exceed about 1 weber per square meter. When the secondary is open, the total flux is produced by the primary current. Writing V_e for the effective value of the primary potential wave, Eq. (12.10) yields

$$\begin{aligned} V_e &= \omega n_p \frac{\phi_m}{\sqrt{2}} \\ &= \frac{2\pi}{\sqrt{2}} \nu n_p A B_m \\ &= 4.44 \nu n_p A B_m \end{aligned}$$

where A is the core area in square meters. Thus, if B_m is to be limited

[1] For a more detailed discussion of this type of diagram see Sec. 13.2.

to unity, the product n_pA must be at least equal to $3.75 \times 10^{-3}V_e$ for a frequency of 60 cycles. For over-all economy and efficiency the core area is generally chosen to be from 1 to 1.5×10^{-4} times the square root of the power to be handled. This provides a general criterion for core design. The length of the magnetic circuit is, of course, kept as short as possible consistent with adequate space for winding and insulation. V_e then determines n_p and n_s is given by the desired output potential. In order to avoid excessive losses in the windings, the current density should not be allowed to exceed 2 or 3 amp. per square millimeter of cross section.

Transformer Connections.—A number of transformer primaries may be connected to the terminals of a phase in either series or parallel.

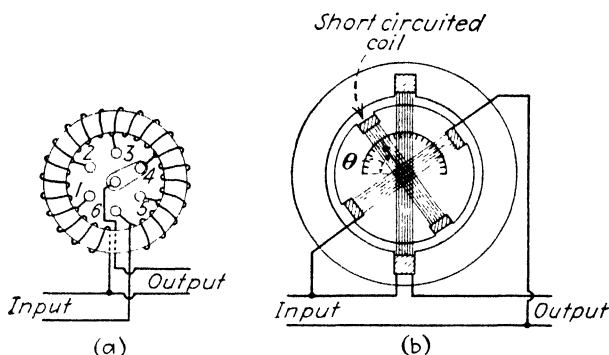



FIG. 12.21.—(*a*) Tapped autotransformer for voltage alteration. (*b*) Induction voltage regulator.

The secondaries may supply separate circuits or may be connected arbitrarily in series to supply the algebraic sum of the induced potentials. The secondaries may also be connected in parallel to increase the current-carrying capacity if the output voltages are the same and the proper sense of connection is observed, *i.e.*, there must be no difference in the potential between terminals connected together. Thus transformer units may be connected in various ways in order to provide the desired output potential and current capacity. A variable output voltage can be obtained from a tapped autotransformer as indicated at the left in Fig. 12.21. A disadvantage of this arrangement is that the secondary circuit is opened when the switch is changed if two taps are not spanned by the arm; if two taps are spanned, a portion of the winding is short-circuited. An alternative arrangement for voltage regulation is to wind one of the coils on a portion of the core that can be rotated with respect to the remainder, as indicated at the right in the figure. The arrangement resembles a motor superficially and there is, of course, a torque between the two coils and the rotating member must be held at the desired angular setting. A short-circuited winding is provided at right

angles to the shunt coil to reduce the reactance of the series winding when the shunt and series windings are in the position of minimum mutual inductance. The mutual inductance between the shunt and series coils is approximately proportional to cos θ and hence the output potential is equal to the input potential plus n_s/n_p cos θ. If $n_s = n_p$, the output potential can be varied from zero to twice that of the input as cos θ changes from -1 through 0 to $+1$. The losses associated with these devices are very small and the control they afford is greatly superior to that of a rheostat.

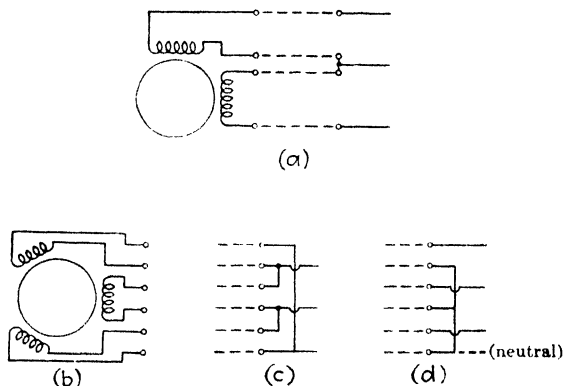



FIG. 12.22.—Polyphase line connections. (*a*) Two phase generator, three-wire power distribution. (*b*) Three-phase generator; (*c*) three-wire distribution Δ; (*d*) three-wire distribution *L*.

In two- and three-phase circuits the phases may be kept entirely separate or they may be combined in various ways for convenience or economy. Thus the windings of a two-phase generator may be connected in series and three wires used for distribution, as shown in Fig. 12.22. In the case of a three-phase generator the windings may all be connected in series if the emfs. are all equal and the proper sense of connection is observed. This is evident from a consideration of the three vectors in Fig. 12.14. The sum of the solid and dashed vector is equal and opposite to the dot-dash vector. Thus if the phases are connected in the proper sequence, there is no potential difference between the last pair of terminals connected. This arrangement is known as the Δ (delta) connection and is indicated at (*c*) in Fig. 12.22. The three junctions are connected to the three-wire distribution line. The potential difference between any pair of lines is the potential developed by the alternator winding. The current flowing in a line is the vector difference of the winding currents; it is equal in magnitude to $\sqrt{3}$ times the winding current as indicated by the transformer diagrams of Fig. 12.23. Alternatively one terminal of each of the three windings may be taken to a common point, called the neutral. The free terminals are connected

to a three-wire transmission line, the neutral may or may not be available. If the sense of connection of the third phase is properly chosen, the phase relation of the emfs. appearing between the lines is that of the potential vectors of Fig. 12.14. This is known as the Y connection. The line potentials are equal to the vector differences of the winding potentials, *i.e.*, $\sqrt{3}$ times them in magnitude, and the line currents are equal to the winding currents.

Representative three-wire three-phase transformer connections are shown in Fig. 12.23. They are of the same nature as the alternator

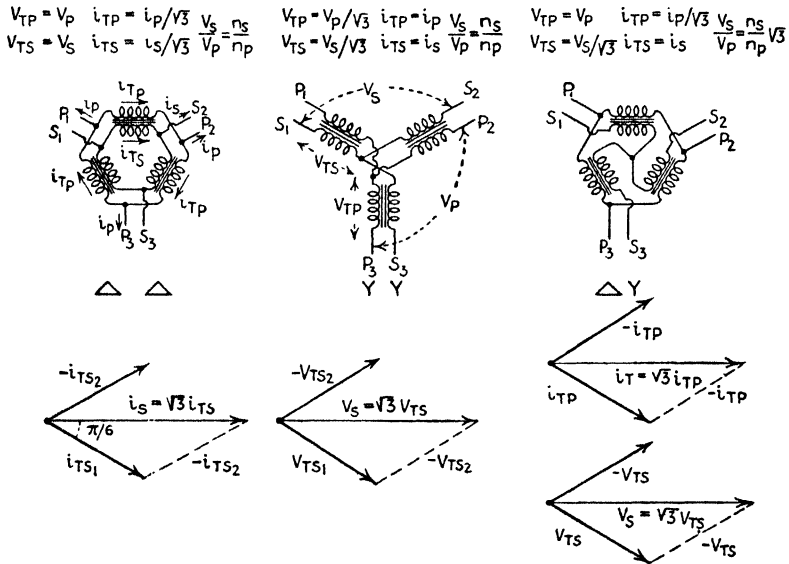



FIG. 12.23.—Representative three-phase transformer connections.

winding connections and are given the obvious designations ΔΔ, YY, and ΔY. In the first type shown at the left the potentials appearing across the transformers are the same as those across the lines, and the line currents are the vector differences of the winding currents, as indicated in the vector diagram. In consequence the windings stand the entire potential but carry only about 58 per cent of the line current. If one transformer is absent, the other two produce the same potential as existed before; thus one of the transformers may fail and be removed without completely disabling the system. The subsequent capacity is smaller by about 42 per cent, but the fact that the system will operate at all is an advantage for this type of connection. The YY connection will not operate as a three-phase system if one transformer is removed. In this type of connection the potential across a winding is only 58 per cent of the line potential, but each winding carries the line current. The neutral may be grounded, which is frequently an advantage. The

ΔY connection is often used for transformers that are designed to increase the voltage. The neutral is grounded and only 58 per cent of the high voltage appears across a transformer winding. If one transformer fails in this connection, it may be removed and the neutral used as the third line, resulting in the same configuration as in the case of the $\Delta\Delta$ connection in which one transformer has failed.

Transformers can also be used for the interconversion of two- and three-phase circuits. Two transformers are required in what is known

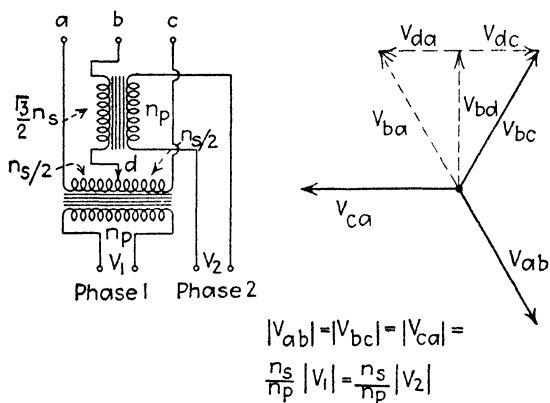



FIG. 12.24.—Scott-connected transformers for the interconversion of two- and three-phase current.

as a *Scott connection.* Each transformer has, say, n_p primary turns. The secondary of one transformer has, say, n_s turns and is center-tapped; the secondary of the other transformer has $\sqrt{3}/2\ n_s$ turns. The connections are indicated in Fig. 12.24 and the illustrative vector diagram is given in the same figure. If V_1 is the potential of one of the two phases, the potential V_{ca} is evidently n_s/n_pV_1. Also

$$V_{ba} = V_{bd} + V_{da} = \frac{\sqrt{3}}{2}\frac{n_s}{n_p}V_2 + \frac{1}{2}\frac{n_s}{n_p}V_1$$

And since V_1 and V_2 are $\pi/2$ apart in phase, V_{ab}, which is equal in magnitude but opposite in sign to V_{ba}, is a vector equal in magnitude to n_s/n_pV_1 and differing in phase by $2\pi/3$. In a similar manner V_{bc} is seen to be a vector of the same length and differing in phase from the other two by $2\pi/3$. The circuit is, of course, reversible and can be used to transform from three phases to two.

12.7. The Rotating Magnetic Field and Induction Motor.—The magnetic field that is produced by an alternating current flowing in a coil of wire may be considered as the sum of two magnetic fields of constant magnitude that rotate in opposite senses with the angular velocity of the alternating current. Assuming linear media, the field or induction

at any point is constant in direction and proportional in magnitude to the current producing it. If i is of the form $i_0 \cos \omega t$, B will be equal to $B_0 \cos \omega t$. The cosine function may be written in terms of the exponential functions as $(e^{j\omega t} + e^{-j\omega t})/2$, and hence

$$\mathbf{B} = \frac{B_0}{2}e^{j\omega t} + \frac{B_0}{2}e^{-j\omega t} \tag{12.14}$$

The vector $e^{j\omega t}$ is a unit vector rotating with an angular velocity ω in the positive sense (counterclockwise) and with the negative sign in the exponent the vector has the opposite sense of rotation. Thus $\mathbf{B}$ may be considered as the sum of the two vectors of magnitude $B_0/2$ rotating in opposite senses. At the center of two circular coils, of the same dimensions but perpendicular to one another with a common diameter, which carry the current from the two phases of a quarter-phase alternator the induction may be considered as a single rotating vector

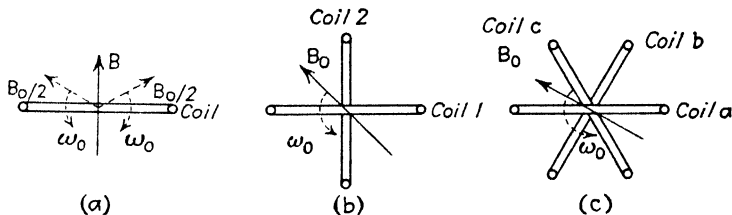



FIG. 12.25.—Rotating magnetic fields. (a) Field due to a single-turn coil may be considered as two constant fields rotating in opposite senses. (b) Rotating magnetic field due to two coils in a two-phase circuit. (c) Rotating magnetic field due to three coils in a three-phase circuit.

of constant magnitude. The geometry is represented in the central diagram of Fig. 12.25. The induction due to one coil is given by Eq. (12.14). The induction due to the other is a similar expression but displaced in phase by the angle $\pi/2$ and in spacial orientation by the same amount. Thus it would be written

$$\begin{aligned}\mathbf{B}_2 &= \frac{B_0}{2}\left(e^{j\left(\omega t+\frac{\pi}{2}\right)} + e^{-j\left(\omega t+\frac{\pi}{2}\right)}\right)e^{\frac{j\pi}{2}} \\ &= \frac{B_0}{2}(e^{j(\omega t+\pi)} + e^{-j\omega t}) \\ &= \frac{B_0}{2}(-e^{j\omega t} + e^{-j\omega t})\end{aligned} \tag{12.15}$$

since $e^{j\pi} = -1$. By geometry the magnitudes of the vectors are equal and on adding Eqs. (12.14) and (12.15) the resultant induction is

$$\mathbf{B} = B_0 e^{-j\omega t} \tag{12.16}$$

This is seen to be a vector of constant magnitude B_0, rotating with the angular velocity ω. The sense of rotation is positive or negative, depend-

ing on the sense of winding and relative orientation of the coils. Similarly, if three coils are disposed as at the right in Fig. 12.25 and carry the current from the windings of a three-phase alternator, a simple rotating magnetic field is produced if the phase and winding sequence is correct. In view of the preceding analysis the induction due to the three coils can be written down from inspection

$$\mathbf{B}_1 = \frac{B_0}{2}(e^{j\omega t} + e^{-j\omega t})$$

$$\mathbf{B}_2 = \frac{B_0}{2}(e^{j\left(\omega t + \frac{4\pi}{3}\right)} + e^{-j\omega t})$$

$$\mathbf{B}_3 = \frac{B_0}{2}(e^{j\left(\omega t + \frac{8\pi}{3}\right)} + e^{-j\omega t})$$

Since the sum of the three unit vectors 1, $e^{\frac{j4\pi}{3}}$, and $e^{\frac{j8\pi}{3}}$ is zero, the resultant induction is

$$\mathbf{B} = \frac{3B_0}{2}e^{-j\omega t} \tag{12.17}$$

This is a vector half again as great as that of the preceding case, rotating in the same sense with the same angular velocity.

The interesting property of a rotating magnetic field is that a piece of metal placed within it experiences a tendency to rotate with the field. The changing induction produces eddy currents that tend to oppose the change and their mechanical reaction drags the metal into rotation with the field. Problem 14 of Chap. X gives the mean torque experienced by a coil which is rotating about a diameter perpendicular to a magnetic field. As it is only the relative motion that is important, a stationary coil would experience the same torque in a rotating magnetic field. If the coil is rotating with an angular velocity ω', the relative rotation $\omega - \omega'$, which written ω_s and called the angular velocity of slip, determines the torque. This is seen to vanish at synchronism, $\omega' = \omega$, and hence, as there is always a retarding frictional torque, the coil will never quite achieve the angular velocity of the rotating field.

This principle is employed in the construction of the induction motor. A three-phase machine is illustrated schematically in Fig. 12.26. The field windings are disposed in the proper sequence around the rotor, though not in general on salient poles as indicated in the figure. The rotor is, of course, of iron to decrease the reluctance of the magnetic circuit. In its periphery are imbedded copper bars which are connected together by copper rings at the ends. These essentially form a rotating secondary circuit. The mechanical reaction of the field, produced by the current in them and the imposed rotating field, gives rise to the motor

torque. As both the magnetic circuit of the field and rotor are subject to a changing induction, they are laminated to reduce eddy-current losses. This type of machine is very efficient and simple in construction and the starting characteristics are generally good. If the rotor is of the wound type in which the armature consists of coils of wire instead of copper bars, the speed and torque can be varied. The terminals of the windings are brought out through slip rings to external variable resistances. It will be seen in the subsequent discussion that a series resistance increases the starting torque and decreases the speed. The use of this device is generally limited to starting, for the presence of resistance in the armature decreases the efficiency. The single-phase

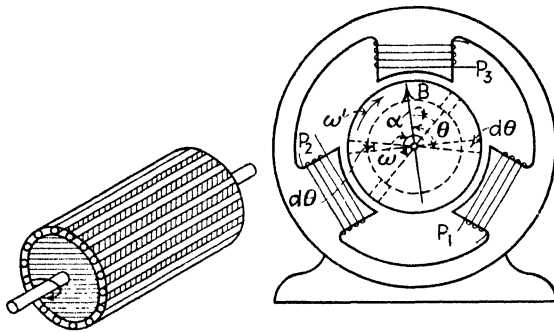



FIG. 12.26.—Rotation of a conducting cylinder in a rotating magnetic field (schematic induction motor).

induction motor is relatively inefficient and this type of construction is limited to small machines. If there is but a single field winding, the motor is evidently not self-starting as the magnetic field consists of two equal vectors rotating in opposite directions. However, if the rotor is given an initial impulse, a steady torque will be developed in that sense as will be seen in the later discussion. An asymmetry that increases the magnitude of one of the rotating vectors at the expense of the other and hence produces a net tendency to rotate in one direction, can be produced by an auxiliary winding in series with a resistance or condenser which alters the phase of the current. Alternatively in very small motors a closed winding may be placed over a fraction of each pole face, accomplishing the same result.

The general principles of operation of an induction motor may be derived quite simply on the basis of certain radical simplifying assumptions. Consider that the rotor is a hollow copper drum mounted in axial bearings and that the magnetic induction is uniform throughout it and rotates with a uniform angular velocity ω. Assume for purposes of analysis that the drum is composed of a large number of rectangular frames disposed at successive angular increments about the axis much

as in the case of the windings of the armature of a direct-current machine. Let the current flowing per unit angle in the sides of the drum be i_θ, where the angle θ is measured from the direction of the magnetic induction **B** which is, of course, continually rotating. The flux through this frame at the angle θ is the flux due to the field, $\phi_m \sin\theta$, plus that due to the currents in the other frames composing the drum. Assume for simplicity that the coefficient of mutual inductance between two frames varies as the cosine of the angle between them, which is approximately correct. Writing L for the coefficient of self-inductance of a single frame, the flux through the frame at an angle θ with **B** due to the current $i_\alpha\, d\alpha$ flowing in a frame at an angle α with **B** is $Li_\alpha \cos(\theta - \alpha)d\alpha$. Thus the total flux through the frame at an angle θ is

$$\phi_\theta = \phi_m \sin\theta + L\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} i_\alpha \cos(\theta - \alpha)d\alpha$$

The emf. induced in the frame is $-\dfrac{d\phi_\theta}{dt}$ which is $-\dfrac{d\phi_\theta}{d\theta}$ times $d\theta/dt$. If the field rotates with the angular velocity ω and the drum with an angular velocity ω', $d\theta/dt = \omega' - \omega = -\omega_s$, where ω_s is the angular velocity of slip. Thus the induced emf. is equal to $\omega_s\, d\phi_\theta/d\theta$, which is also equal to the current i_θ times the resistance R of the frame, or

$$\mathcal{E}_\theta = Ri_\theta = \phi_m\omega_s \cos\theta - L\omega_s\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} i_\alpha \sin(\theta - \alpha)d\alpha \qquad (12.18)$$

This is an integral equation for i_θ. Assume a solution

$$i_\theta = i_m \cos(\theta + \beta) \qquad (12.19)$$

Inserting this expression as $i_\alpha = i_m \cos(\alpha + \beta)$ in the last term and performing the integration yields $-\dfrac{L\omega_s i_m\pi}{2}\sin(\theta + \beta)$. Substituting the value of i_θ on the left of the equality sign also and expanding the trigonometric sums

$$Ri_m \cos\beta\cos\theta - Ri_m \sin\beta\sin\theta = \phi_m\omega_s\cos\theta - L\omega_s i_m\frac{\pi}{2}\cos\beta\sin\theta - L\omega_s i_m\frac{\pi}{2}\sin\beta\cos\theta$$

If this equation is to be true for all values of θ, the coefficients of the sine and cosine functions of θ on the two sides must be equal, which yields the two conditions determining i_m and β as

$$i_m = \phi_m\omega_s\left[R^2 + \left(L\omega_s\frac{\pi}{2}\right)^2\right]^{-\frac{1}{2}} \qquad \text{and} \qquad \beta = \tan^{-1}\left(L\omega_s\frac{\pi}{2R}\right) \qquad (12.20)$$

The torque on a frame at an angle θ with the flux and carrying a current $i_\theta\, d\theta$ is $\phi_m \cos\theta\, i_\theta\, d\theta$. Thus from Eqs. (12.19) and (12.20) the torque on the entire rotor is given by

$$\begin{aligned} T &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \phi_m \cos\theta\, i_\theta\, d\theta \\ &= \phi_m i_m \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos\theta \cos(\theta + \beta)\, d\theta \\ &= \phi_m i_m \frac{1}{2} \cos\beta \\ &= \frac{\pi \phi_m^2 \omega_s R}{2[R^2 + (L\omega_s \pi/2)^2]} \end{aligned} \tag{12.21}$$

This expression is plotted as a function of ω' in Fig. 12.27. When the rotor is at rest, $\omega' = 0$ and $\omega_s = \omega$. The torque is not large, but it is in the sense of ω. As ω' increases, the denominator decreases more rapidly than the numerator and the torque increases; thus in the case of a single-phase machine a net torque develops in the sense of initial rotation. The torque reaches a maximum when

$$\omega_s = \frac{2R}{\pi L} \tag{12.22}$$

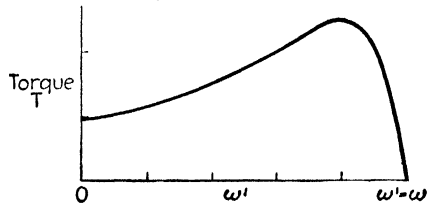



FIG. 12.27.—Torque on the armature of an induction motor as a function of the speed.

as may be seen by differentiation of Eq. (12.21). Since ω_s is large at starting, the starting torque is increased by increasing the rotor resistance. This may be accomplished for a wound rotor by a series resistance which is cut out when the machine gains speed. In practice the machine is used over the portion of the curve to the right of the maximum where the slip is small. Here the second term in the denominator is negligible to a first approximation and the torque is given by $T = \pi\phi_m^2\omega_s/2R$. Thus the torque is proportional to the slip and inversely proportional to the rotor resistance. As $\dfrac{d\omega'}{dT} = -\dfrac{2R}{\pi\phi_m^2}$, the speed of rotation decreases with increased load, but only slightly if R is small. In practice the angular velocity of slip is only a few radians per second and the motor rotates at almost the synchronous speed.

Alternatively the induction motor can be considered as a rotary transformer with a mechanical output, which it essentially is. The losses in general are not negligible, though they are exaggerated for purposes of clarity in the vector diagram of Fig. 12.28. When the motor is run-

ning freely, the effective impedance is very large, the back emf. practically equals the applied potential, and the only current flowing is the small one producing the magnetization. As the load and hence the slip are increased, power is delivered to the rotor secondary and to the load, as would be the case if the resistance in the secondary of a transformer were decreased. The schematic circuit is indicated at the top in Fig. 12.28. Linear circuit parameters are assumed. The magnetizing current i_m flowing in the dashed circuit accounts for the core losses. The effective ohmic losses in the primary and secondary circuits occur in the lumped resistance R, and X represents the effective leakage reactance. These are in general small in comparison with the load resistance R_l. The power consumption in this represents the mechanical power delivered by the motor.

FIG. 12.28.—(a) Schematic circuit of an induction motor. (b) Vector diagram of an induction motor considered as a rotating transformer.

The accompanying vector diagram illustrates the analysis of this circuit. The vector V represents the applied potential and ϕ, which is the magnetic flux, lags in phase by $\pi/2$. i_m, which is small and leads the flux by a small angle, is the approximately constant magnetizing current flowing through the shunt circuit. The phase and magnitude of the current i' are determined by R, R_l, and X. As was seen in the transformer discussion of Sec. 12.6, the locus of the tip of the vector i' as the circuit resistance is varied is a circle of diameter V/X, as indicated in the figure. The distance AG of the figure is equal to $V/2X$. The total current i is the sum of i_m and i'; its locus is, of course, the same circle, and the cosine of the angle between i and V is the power factor of the motor. The product of V and the projection of i upon it is the total power consumed by the motor and the product of V and the projection of i' upon it is the power consumed exclusive of the magnetizing losses, *i.e.*, the power delivered to the series circuit of R, R_l, and X.

The data for the construction of the circle diagram can be obtained by calculation from the design or by measuring the magnitude and phase angle of the current at no load and with the rotor clamped. In the latter

measurement the line voltage is reduced till the motor draws a normal current; the actual stalled current that would be drawn is obtained by multiplying by the appropriate ratio. These measurements give, say, the currents $i_m = i_{OA}$ and i_{OF}. The vector i_{AF} is then a chord of the circle. Thus, if the perpendicular bisector is erected, its intersection with the line through A parallel to ϕ is the center of the circle. The product of V and i_{DF} is the power delivered to the stalled motor, *i.e.*, to the resistance R. If the ratio of the primary resistance to the secondary resistance is DE/EF, the product i_{DE} and V is the stalled power delivered as heat to the primary and Vi_{EF} is the same quantity for the secondary. The proportion of R associated with the primary and secondary must be determined by a separate measurement if the losses are to be attributed to the different circuits. Since the same current flows through all the resistances, R which is the sum of the primary and secondary resistances and R_l the load, the power delivered to each is proportional to that resistance. Thus the power dissipation in the separate resistances can be determined graphically for any arbitrary current, say, that corresponding to the point P on the circle. Vi_{BP} is the power delivered to the load, *i.e.*, the product $T\omega'$, where T is the torque and ω' is the angular velocity of rotation. Vi_{CB} is the heating loss in the secondary and Vi_{HC} is the heating loss in the primary. Vi_{JH} is the core loss. The efficiency is the ratio of the lengths BP and JP. The maximum output occurs for the maximum value of i_{BP} along the circle. Evidently the point on the circle corresponding to the maximum output is determined by the point of tangency of a line parallel to AF. The point P chosen in the diagram corresponds to the maximum output.

It should also be mentioned that it is possible to operate the direct-current commutator type of motor on an alternating-current line. It was seen in the discussion of this type of motor that the torque is in the same direction for either sense of the applied potential. Hence the torque continues in the same direction for both halves of the alternating-current wave, and the motor can be operated from an alternating-current line. There are certain essential modifications, however, in the design of an alternating-current commutator motor. In the first place, both the armature and field carry an alternating current and the entire iron path, which includes that through the field, must be laminated to reduce eddy-current losses. Furthermore, the self-inductance of the armature is very large and its effect must be reduced in some way in order that it shall not limit the flow of current through the windings. This is generally accomplished by auxiliary field windings that operate on the transformer principle. They are disposed physically at right angles to the principal field in the manner of the direct-current interpoles of Fig. 12.7. They may be connected in series with the armature in such a sense as to oppose the induction resulting from the current in the rotor. This is analogous to two transformer windings connected in series in opposing senses in which case the net self-inductance is negligible. Alternatively the auxiliary windings may form an entirely separate closed circuit as in the case of the short-circuited winding on the induction-voltage regulator of Fig. 12.21. In this connection they act as a short-circuited transformer secondary. If the

impedance of the secondary circuit is vanishingly small, that of the primary (which is the armature) is also small by Eq. (12.13) and the desired result is accomplished. The commutator type of alternating-current motor generally has a series characteristic and is widely used for variable-speed operation.

Problems

1. Derive Eq. (12.3) by applying the general force equation [Eq. (9.6)] to a drum armature.

2. The drum armature of a two-pole generator has 12 coils of 20 turns apiece. Its effective length is 0.167 m. and its effective radius is 0.05 m. If it rotates in a region of uniform induction of 1 weber per square meter at a rate of 1,800 r.p.m., show that the emf. induced is 120 volts. Taking the rated current as that which gives a product of the current times the number of conductors per centimeter of armature periphery of 125, show that the rated current output is 8.2 amp. and that the rated power output is 1 kw. Assuming a 90 per cent efficiency, what must be the horsepower of the machine driving it?

3. Assume a machine to which the magnetization curve of Fig. 12.8 applies and that a current of 0.6 amp. corresponds to an induction of 1 weber per square meter. If the reluctance of the magnetic circuit is essentially that of the air gap between the field and armature with an effective length of 0.5 cm., show that the field coils must have 6,670 turns.

4. If the resistance of the field coils of the preceding problem is 125 ohms and the armature resistance is 1 ohm and the machine is used as a shunt generator, derive graphically the output characteristic (output voltage as a function of output current). What would be the maximum output of the machine?

5. A homopolar generator of the type of Fig. 12.11 consists of a steel disk 30 cm. in radius rotating at the rate of 3,000 r.p.m. Assuming that the field coils have 1,000 turns and the reluctance of the magnetic circuit is that of the air gap, which is effectively 1 cm. long, show that the field coils would have to carry a current of 8 amp. in order that the disk may generate an emf. of 14 volts. What is then the flux density in the gap? If the disk and contacts have a resistance of 0.098 ohms, show that the maximum output is 1 kw.

6. The homopolar generator of the preceding problem is run as a separately excited motor with an induction of 1 weber per square meter. If a potential difference of 16.2 volts is applied to the disk contacts and a current of 50 amp. flows through them, show that the disk rotates at 2,400 r.p.m. and exerts a torque of 0.23 kg.-m. Find the efficiency of the armature circuit.

7. A 10-hp. direct-current shunt motor of 90 per cent efficiency is connected to a 240-volt line. If the field current is 1 amp. and the armature resistance is 0.2 ohms, show that: power input = 8.3 kw., armature current = 33.5 amp., back emf. = 233.3 volts. If the speed of rotation is 600 r.p.m., show that the torque exerted is 12.1 kg.-m.

8. Show that in the case of the motor of the preceding problem a starting resistance of 6.95 ohms would have to be inserted in series with the armature in order to limit the starting current to its full-load value. What would then be the starting torque? If this series resistance is reduced to 1 ohm and the motor operates at the full-load torque, show that the speed is reduced to 535 r.p.m. and the efficiency to 80 per cent.

9. The reluctance of the magnetic circuit of a series motor is 2×10^5 amp.-turns per weber. If the field has 100 turns and the armature 1,000 and a stalled or starting current of 30 amp. flows, show that the torque is 14.6 kg.-m. Assuming a 240-volt line and a field resistance of 1 ohm, show that the motor runs at 570 r.p.m. and draws a current of 22.8 amp. if it is operating at 80 per cent efficiency against half the starting torque. What is then the power output of the machine?

10. At what frequency must a 24-pole alternator rotate to generate a 60-cycle frequency? If the coils of a phase are in series and have 20 turns apiece, show that the maximum flux through a coil must achieve the value 0.0343 webers to generate an effective emf. of 2,200 volts.

11. Assuming a sinusoidal distribution of the induction about the periphery of the armature and a maximum value of 1 weber per square meter, show that the inner radius of the armature of the machine of the preceding problem must be about 0.83 m. if the width of the armature parallel to the shaft is 0.25 m.

12. Two six-pole alternators on the same shaft have their outputs connected in series. If the rotating member of one is 20 mechanical degrees ahead of the other, show that the total voltage developed is either the same or $\sqrt{3}$ times as great, depending on the sense of the series connection.

13. The excitation is such that the power factor of a synchronous motor is unity. Assuming a 10-kw. output and a line voltage of 240, show that the emf. that would be generated by the motor run backward at the same speed is 254 volts if the reactance of the windings is 2 ohms at the applied frequency.

14. A 10-kw. 60-cycle transformer is designed to operate from 2,400 to 120 volts. Show that the core area should be about 120 cm.^{2} and if the flux density is not to rise above 1 weber per square meter the primary should have 750 turns at this core area. Find the proper number of secondary turns, assuming negligible losses, and show that the total effective conductor cross section of the two windings should be about 30 cm.^{2}.

15. A 10,000-volt generator supplies 100 kw. to a 100-ohm transmission line. What must be the transformation ratio of a terminating autotransformer in order to bring the output voltage up to that of the generator? Assuming a transformer efficiency of 97 per cent, what is the available power output? (Assume unit power factor.) Need the entire winding of the autotransformer have the same current-carrying capacity?

16. What must be the turn ratio of an induction-voltage regulator to vary the output voltage by plus or minus 30 per cent? Plot the input-output voltage ratio as a function of the angular setting of the regulator.

17. Three equal resistances are connected in Δ across the terminals of a three-phase line. Show that the heat developed is three times as great as if they were connected in Y. If the line potential is 120 volts and the resistances are 10 ohms apiece, show that the rate at which heat is developed in the two cases is 4.32 and 1.44 kw., respectively.

18. Show that if 51 lamps can be lit to rated brilliance (terminal voltage) when connected in Δ to three-phase mains, only 30 lamps can be lit to the same brilliance when connected in Y. Show that the power consumed in the second case is only 58 per cent of that in the first.

19. Show that the total power delivered by a three-phase line to a symmetrical load is $\sqrt{3}\ V_e i_e \cos \alpha$, where V_e is the effective voltage between lines, i_e is the effective current flowing in them, and α is the phase angle of the load.

20. Show that the current flowing in each three-phase line of Scott-connected transformers is greater by the ratio $2n_p/\sqrt{3}n_s$ than that in each of the two-phase lines. It is desired to supply 120-volt three-phase power to a 10-kw. load from a 2,400-volt two-phase line. Specify the appropriate core areas, numbers of turns, and the cross section of the wires for the two transformers.

21. The sense of winding of one of the three coils giving rise to a rotating magnetic field is reversed. Show that the resulting magnetic field can be considered as the alternating field produced by the reversed coil plus a field one-third as great as the original one and rotating in the opposite sense.

22. A 110-volt induction motor draws a current of 1 amp. at a phase lag of $2\pi/5$ when running freely. If the rotor were clamped, it would draw a current of 25 amp. at the same power factor. Plot the operating circle diagram. Show that the effective reactance at the applied frequency is 4 ohms. Show that the maximum output is 850 watts. Show that the efficiency and power factor at maximum output are 73 and 76 per cent, respectively.

CHAPTER XIII

SIMPLE CIRCUITS CONTAINING INDUCTANCE, CAPACITANCE, AND RESISTANCE

13.1. Free Oscillations.—Introductory discussions of resistance-capacity and resistance-inductance circuits were given in Secs. 7.6 and 10.1. A more complete analysis has been postponed until the general case of a circuit containing all three types of elements could be considered. The differential equations for the charge q [Eq. (7.12)] and for the current i [Eq. (10.2)] were both first-order linear equations with constant coefficients, and hence the solutions were of the same type. [For the general solution see Eq. (C.6) of the Appendix.] When, however, the circuit contains both capacity and inductance, the differential equation for either the charge on the condenser or the current through the circuit is of the second order and the physical phenomena represented by such an equation are of an entirely different type. Consider first the circuit represented by Fig. 13.1. The general circuital law states that the sum of the potential differences between the terminals of all the elements of a circuit (with due regard to sign) is equal to zero. On applying the law to this circuit

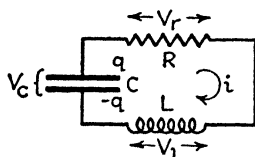



Fig. 13.1.—Circuit containing inductance, capacitance, and resistance.

$$V_c = \frac{q}{C}$$

$$V_R = iR = R\frac{dq}{dt}$$

$$V_L = L\frac{di}{dt} = L\frac{d^2q}{dt^2}$$

or

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q = 0 \qquad (13.1)$$

Differentiation by t shows that i obeys an exactly similar equation, as do the separate potentials V_c, V_R, and V_L. It is recognized as the differential equation of damped simple harmonic motion which is very familiar in mechanics. In mechanical motion L represents the reaction per unit acceleration (the mass or moment of inertia), R is the retarding force per unit velocity, and $1/C$ is the restoring force per unit displace-

ment. The similarity of the equations implies a superficial similarity in the nature of the observed phenomena. For the same conditions determining the two arbitrary constants contained in the solution, the charge on the condenser is the same function of the time as is the displacement of the mass in a mechanical system. Similarly i corresponds to the velocity and di/dt to the acceleration. Mechanical systems of this type such as springs, pendulums, etc., are familiar and the analogy is very useful in visualizing the behavior of the general L-R-C circuit. The displacement of a galvanometer movement obeys this equation, and one periodic solution [Eq. (10.10)] has already been considered in detail.

This equation [Eq. (C.2)] and its general solutions [Eqs. (C.12), (C.13), and (C.15)] are considered in the Appendix. As an example of the application of these to the circuit of Fig. 13.1 consider the particular case in which the condenser is initially charged to a potential q_0/C and a switch in the circuit then closed. If t is measured from the closing of the switch, $q = q_0$ and $i = 0$ at $t = 0$. In terms of the coefficients of Eq. (13.1) the solutions given in the Appendix become

$$q = q_0\left(\frac{R^2C}{4L} - 1\right)^{-\frac{1}{2}} e^{-\frac{Rt}{2L}} \sinh\left[\left(\frac{R^2}{4L^2} - \frac{1}{LC}\right)^{\frac{1}{2}} t + \delta\right]$$

where

$$\delta = \tanh^{-1}\left(1 - \frac{4L}{R^2C}\right)^{\frac{1}{2}} \qquad \text{if} \qquad R^2 > \frac{4L}{C}$$

$$q = q_0\left(1 + \frac{Rt}{2L}\right)e^{-\frac{Rt}{2L}} \qquad \text{if} \qquad R^2 = \frac{4L}{C}$$

$$q = q_0\left(1 - \frac{R^2C}{4L}\right)^{-\frac{1}{2}} e^{-\frac{Rt}{2L}} \sin\left[\left(\frac{1}{LC} - \frac{R^2}{4L^2}\right)^{\frac{1}{2}} t + \delta'\right]$$

where

$$\delta' = \tan^{-1}\left(\frac{4L}{R^2C} - 1\right)^{\frac{1}{2}} \qquad \text{if} \qquad R^2 < \frac{4L}{C}$$

The first case is the aperiodic solution for which there is only one maximum value of q which occurs at the origin ($t = 0$). The hyperbolic sine function approaches one-half the exponential function for large values of the argument; hence after a certain time q decreases approximately exponentially. The current in the circuit can be obtained by taking the derivative of q with respect to t. Its maximum occurs after $t = 0$, and later it also decreases exponentially. The second case is the limiting one of critical damping. It is also approximately exponential for sufficiently great times. The rate of decrease of q is greater than for any of the aperiodic cases. The third case represents the periodic type of solution in which q is given by a sine function with an expo-

nentially decreasing amplitude. This is the type of solution that was discussed for the galvanometer movement in Sec. 10.5. The charge on the condenser is plotted as a function of t for the three cases in Fig. 13.2. Representative values of the parameters are assumed and the nature of the variation of the current with the time is also indicated.

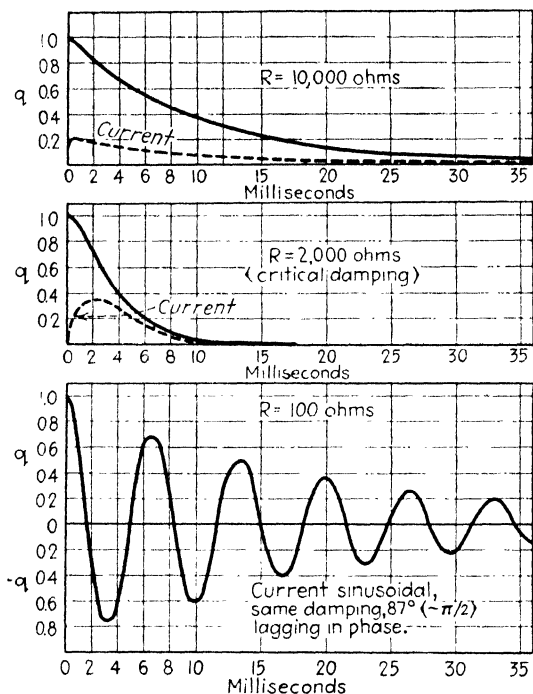



FIG. 13.2.—Discharge of a 1 microfarad condenser through a 1 henry inductance for three circuit resistances.

The periodic case is the one of greatest interest in general and it presents certain features which have not been previously encountered in the discussion of circuit theory. In the limiting case of a negligible resistance the charge on the condenser is given by

$$q = q_0 \sin\left(\omega_0 t + \frac{\pi}{2}\right) = q_0 \cos \omega_0 t$$

where ω_0 is written for $1/\sqrt{LC}$, and the current is given by

$$i = i_0 \sin \omega_0 t$$

where $i_0 = -\omega_0 q_0$. These quantities are both periodic with a period $\tau_0 = 2\pi/\omega_0$ and they are $\pi/2$ out of phase. The energy of the system is that which was originally put into the condenser, namely, $\frac{1}{2}q_0^2/C$. Since both the capacity and inductance are assumed resistanceless, no

energy is dissipated and the sum of the electrical energy in the condenser and the magnetic energy in the inductance is constant, *i.e.*,

$$\frac{1}{2}\frac{q^2}{C} + \frac{1}{2}Li^2 = \frac{1}{2}\frac{q_0^2}{C} = \frac{1}{2}Li_0^2$$

This may also be seen by forming the equation from the previous expressions for i and q. The general case in which resistance is present is a little more involved. Both q and i are periodic with the period $\tau = 2\pi/\omega$, where ω is written for $\left(\omega_0^2 - \frac{R^2}{4L^2}\right)^{1/2}$ which is the coefficient of t in the argument of the sine function. It is seen from this that the period τ is greater than the free period τ_0. The ratio of successive maxima of either function, which occur at the interval τ, are seen to be $e^{-\frac{R\tau}{2L}}$. The exponent without the minus sign is known as the logarithmic decrement δ; thus $\delta = R\tau/2L$.

It is very convenient in much of the subsequent discussion to introduce a quantity Q which is a figure of merit for an oscillatory circuit. It is defined for later work as the ratio of the inductive reactance of a circuit to its resistance, *i.e.*,

$$Q = \frac{\omega L}{R}$$

On multiplying numerator and denominator by $i_0^2/2$ the numerator is seen to be ω times the maximum energy stored in the inductance; *i.e.*, the energy of the circuit oscillation and the denominator, which is also i_e^2R, is the rate at which this energy is being lost from the oscillatory form and converted into heat. Thus the definition of Q is equivalent to

$$Q = \frac{\text{(circulating energy)}}{\text{(rate of loss of circulating energy)}}$$

or

$$Q = 2\pi\frac{\text{(circulating energy)}}{\text{(energy loss per cycle)}}$$

For free oscillation ω is the natural angular frequency of the circuit and $\delta = \pi/Q$. In later work ω will be a variable and here it is a rather complicated function of the circuit parameters; hence an auxiliary quantity Q_0 is useful for periodic circuits. It is defined as

$$Q_0 = \frac{\omega_0 L}{R} = \frac{1}{R}\sqrt{\frac{L}{C}}$$

The decrement can be written in terms of this quantity and the period

in terms of it and ω_0

$$\delta = \frac{\pi}{Q_0}\left[1 - \left(\frac{1}{2Q_0}\right)^2\right]^{-\frac{1}{2}}$$
$$\tau = \frac{2\pi}{\omega_0}\left[1 - \left(\frac{1}{2Q_0}\right)^2\right]^{-\frac{1}{2}}$$

An efficient periodic circuit is one that has a small resistance or a large Q_0. If Q_0 is very much greater than unity, δ is approximately equal to π/Q_0 and τ is approximately equal to $2\pi/\omega_0$. It is evident from the definition that Q_0 must be greater than one-half for the circuit to be naturally periodic. If the circuit is to be kept in sinusoidal oscillation with an effective current i_e, power must be supplied to it at the rate i_e^2R, for this is the rate of dissipation. Thus, if P is the rate at which power is supplied to the circuit, the effective circulating current is given by

$$i_e = \sqrt{\frac{P}{R}}$$

For continuous oscillation the mean energy in the electric and magnetic form are equal, $\frac{1}{2}CV_{ce}^2 = \frac{1}{2}Li_e^2$, where V_{ce} is the effective potential difference across the condenser. Hence, in terms of the rate of supply of power to the circuit

$$V_{ce} = \sqrt{\frac{Q_0P}{\omega_0C}} = \sqrt{\frac{LP}{RC}}$$

The maximum potential occurring across the condenser is, of course, $\sqrt{2}$ times V_{ce}. These equations are frequently very useful in determining circuit parameters, particularly at high frequencies. If the current and rate of supply of power are measured, R can be obtained. C is generally calculable, and a measurement of ω_0 determines L. If, however, C is not known, measurements of ω_0 and V_{ce} determine both L and C.

13.2. Forced Oscillations in a Series Circuit.—A battery of constant potential V can be introduced in the circuit of Fig. 13.1 without changing the nature of the observed phenomena. Equation (13.1) then has a constant V on the right-hand side, but the variable can be changed to $q' = q - VC$, which returns the equation to its original form. As nothing new is involved, this circuit will not be considered further. The most general case of the simple linear circuit is that in which it contains an applied potential which is a function of the time, $V(t)$. The solution of the differential equation representing the circuit in this case is given in the Appendix [Eq. (C.16)]. For a discussion of the transients and the other important phenomena associated with special

forms of $V(t)$ the reader is referred to treatises on circuit theory.[1] Our attention will be confined to the so-called *steady-state solutions* which represent the current and the potential differences across circuit elements a considerable time after the conditions have been established. The special discussion will be in terms of a simple periodic potential. The general alternating potential can be expressed in a Fourier series as a sum of simple periodic ones with frequencies that are integral multiples of the fundamental frequency. As the differential equation is linear, the solution for the sum is simply the sum of the solutions for the potentials separately with due regard to amplitude, frequency, and phase. Hence it is only necessary to consider the nature of the current for one simple periodic electromotive force.

Let us assume that the potential difference applied to the circuit by some type of generator is $V = V_0 \cos \omega t$, which is the same as the real part of

$$\mathbf{V} = V_0 e^{j\omega t}$$

With this inclusion Eq. (13.1) has V on the right-hand side. Differentiating with respect to t to obtain the equation in terms of the current as the variable, and employing the complex notation with the understanding that the real part of the solution is the actual current, the equation becomes

$$L\frac{d^2\mathbf{i}}{dt^2} + R\frac{d\mathbf{i}}{dt} + \frac{\mathbf{i}}{C} = j\omega\mathbf{V} = V_0 j\omega e^{j\omega t} \tag{13.2}$$

The current in the steady state is also simply periodic and we may assume the solution $\mathbf{i} = i_0 e^{j\omega t}$. Substituting this in the equation

$$\left(-L\omega^2 + Rj\omega + \frac{1}{C}\right) i_0 e^{j\omega t} = j\omega V_0 e^{j\omega t}$$

Dividing through by $j\omega$, the complex current is given by

$$\mathbf{i} = \frac{\mathbf{V}}{\mathbf{z}} \tag{13.3}$$

where $\mathbf{z} = R + j\left(\omega L - \frac{1}{\omega C}\right)$ is the complex impedance.[2] If $\mathbf{z}$ is written in terms of its magnitude, $Z = \left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2}$, and its phase

[1] BUSH, "Operational Circuit Analysis," John Wiley & Sons, Inc., New York, 1929; CARSON, "Electric Circuit Theory," McGraw-Hill Book Company, Inc., New York, 1926; GARDNER and BARNES, "Transients in Linear-Systems," John Wiley & Sons, Inc., New York, 1942.

[2] The reciprocal of the complex impedance is known as the *admittance* $\mathbf{y}$; $\mathbf{y} = 1/\mathbf{z}$. The real and imaginary parts of $\mathbf{y}$ are known as the *conductance* g and the *susceptance* b, respectively.

$$\text{angle } \varphi = \tan^{-1} \frac{\left(\omega L - \dfrac{1}{\omega C}\right)}{R}$$

$$\mathbf{i} = \frac{V_0}{Z} e^{j(\omega t - \varphi)} \tag{13.4}$$

In terms of the actual potential and current, which are the real parts of these complex vectors

$$V = V_0 \cos \omega t$$
$$i = \frac{V_0}{Z} \cos (\omega t - \varphi) \tag{13.5}$$

As in the vector diagrams of Secs. 7.6 and 10.1, these are the projections on the real axis of the vectors **i** and **V**, which may be thought of as rigidly attached together and rotating about the origin in a counterclockwise sense (that of positive angle) with an angular velocity ω. The magnitude of the current vector is $i_0 = V_0/Z$, and the effective values of the vectors are, of course, also in the same ratio. The current lags behind the potential by the angle φ, and the power factor of the circuit which is $\cos \varphi$ is

$$\text{Power factor} = \cos \varphi = \frac{R}{\left[R^2 + \left(\omega L - \dfrac{1}{\omega C}\right)^2\right]^{\frac{1}{2}}} = \frac{R}{Z} \tag{13.6}$$

The instantaneous consumption of power, iV, is the rate at which energy is stored in the condenser and inductance as well as that at which it is dissipated in the resistance. The mean consumption of power by the circuit is the average value of the product of Eqs. (13.5) over a complete period which by Sec. 5.5 is seen to be

$$P = \frac{V_0^2}{2Z} \cos \varphi = \frac{V_e^2}{Z} \cos \varphi = V_e i_e \cos \varphi = i_e^2 R \tag{13.7}$$

This analysis is formally identical with that of Secs. 7.6 and 10.1, but the actual phenomena observed on variation of the parameters or of the frequency present a somewhat different aspect. The vectorial method is the most instructive for considering the variation of the circuit parameters. From the previous discussion the magnitude of the current, i_0, can be written in either of two forms

$$i_0 = \frac{V_0}{R} \frac{R}{Z} = \frac{V_0}{R} \cos \varphi$$

or

$$i_0 = \frac{V_0}{X} \frac{X}{Z} = \frac{V_0}{X} \sin \varphi$$

where X is the reactance. Now the equation $r = a \cos \varphi$ is the polar equation of a circle of diameter a passing through the origin and tangent to the y or imaginary axis, and $r = a \sin \varphi$ is the equation of a circle through the origin tangent to the x or real axis. Thus, if R is a constant and the reactance is varied, the locus of the tip of the current vector is a circle with its center a distance $V_0/2R$ along the positive real axis and a diameter V_0/R. This is known as a *circle diagram*. If the impedance is constant and R is varied, the second form shows that the locus of the tip of the current vector is a circle of diameter V_0/X with its center a distance $V_0/2X$ along the imaginary axis. The fact that the param-

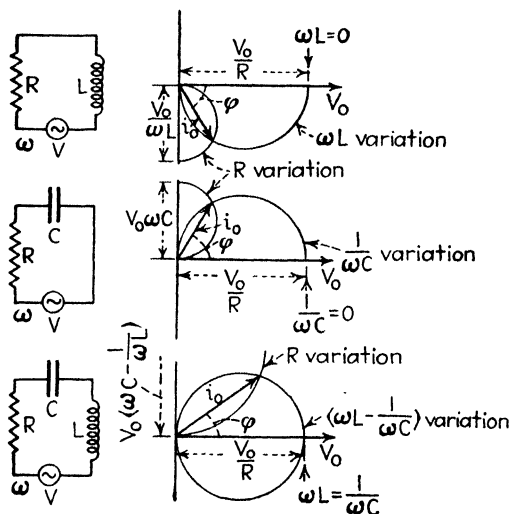



FIG. 13.3.—Complex vector analysis of simple alternating-current circuits.

eters cannot all assume either positive or negative values introduces a certain asymmetry. Circle diagrams for R-L, R-C, and R-L-C circuits are indicated in Fig. 13.3. When only inductance is present, both R and X (ωL) are limited to positive values, and only the semicircles lying in the lower right-hand quadrant can actually be described. When only capacity is present, R is positive and $X\left(-\dfrac{1}{\omega C}\right)$ is negative and the available semicircles are those in the upper right-hand quadrant. When all three elements are present, X can be either positive or negative, though R is, of course, still limited to positive values and both right-hand quadrants are available. The actual current vector is, of course, that drawn from the origin to the intersection of the two circles specified by the values of the parameters. Its variation as one of the parameters is varied is brought out by considering variations in one of the circle diameters. Since ω is involved in X, these circles also indicate the

variation of i with frequency. The circle diagram is a very useful device for bringing out the general nature of the dependence of i on the circuit parameters and the frequency in all different types of alternating-current circuits and it will be referred to frequently in this and the following chapter.

From either the circle diagram or the equations it is evident that the current is a maximum and in phase with $\mathbf{V}$ for R-C and R-L circuits only for extreme values of the parameters, *i.e.*, $\omega L = 0$ or $1/\omega C = 0$. The complete L-R-C circuit differs in this respect for the current is a maximum and φ is equal to zero for

$$\omega L = \frac{1}{\omega C} \qquad \text{or} \qquad \omega = \frac{1}{\sqrt{LC}} = \omega_0 \tag{13.8}$$

The condition for which $\mathbf{i}$ is in phase with $\mathbf{V}$ is known as *resonance*. The resonant frequency of a simple series circuit is the natural frequency of the circuit for zero resistance. And in this case the resonant condition also corresponds to that of minimum impedance or maximum current. It is the phenomenon of resonance that is the particular characteristic of the L-R-C circuit. The circle diagram shows qualitatively the way in which the magnitude of the current varies with the frequency and the fact that it leads the potential for $\omega < \omega_0$ and lags behind it for $\omega > \omega_0$. The details are more clearly brought out by the analytic expressions for Z and φ. In terms of Q_0 and ω_1 these are

$$\frac{Z}{R} = \left[1 + Q_0^2\omega_1^2\left(1 - \frac{1}{\omega_1^2}\right)^2\right]^{\frac{1}{2}}$$

and

$$\varphi = \tan^{-1} Q_0\omega_1\left(1 - \frac{1}{\omega_1^2}\right)$$

where

$$Q_0 = \frac{1}{R}\sqrt{\frac{L}{C}} \qquad \text{and} \qquad \omega_1 = \frac{\omega}{\omega_0}$$

As the magnitude of the current vector is proportional to R/Z, this quantity is plotted as a function of ω_1 in Fig. 13.4 for three representative values of the parameter Q_0. The maximum value of i_0 occurs at $\omega_1 = 1$ for all values of Q_0 and is of course equal to V_0/R. The curves are not symmetrical on either side of the maximum, but they are approximately so in its immediate neighborhood. The phase angle φ is not plotted, but it can be seen from its dependence on ω_1 to vary from $-\frac{\pi}{2}$ for very small ω_1's through 0 for $\omega_1 = 1$ to $\pi/2$ for very large values of ω_1.

Circuits for which Q_0 is large find many important applications. These are dependent largely on the rapid change in impedance in the

neighborhood of $\omega_1 = 1$. If $\Delta(\omega)$ represents a very small percentage change in frequency from ω_0, the percentage change in current $\Delta(i)$ that results is given approximately by

$$\Delta(i) = -2Q_0^2[\Delta(\omega)]^2$$

Thus, if Q_0 is of the order of 250, a 0.1 per cent change in frequency will result in a 12.5 per cent change in current. It is evident that Q_0 is a measure of the discrimination of the circuit, *i.e.*, the change in current

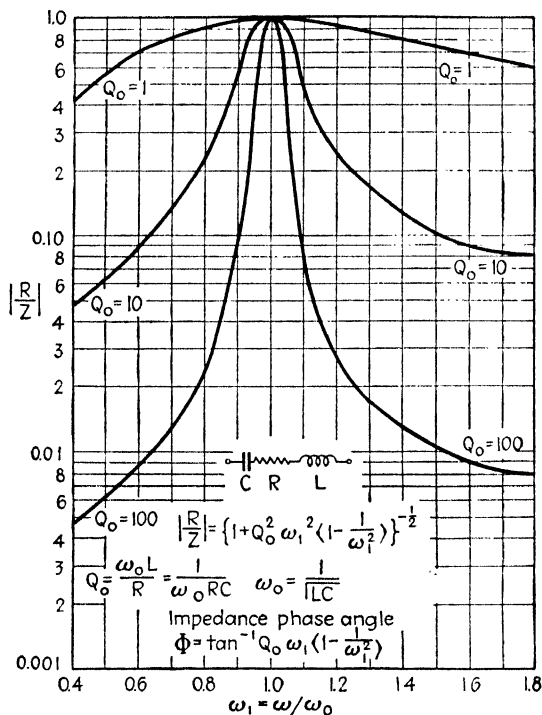



FIG. 13.4.—Curves representing the variation of the current flowing through a series L-R-C circuit as a function of the frequency for representative values of Q_0.

that will occur for a certain change in frequency. One application of this differential response with frequency is in the *wavemeter* or *frequency meter*. In its simplest form this is merely a coil of wire with a condenser across its terminals. The mutual inductance between a neighboring circuit and this coil induces a current in the wavemeter circuit. The magnitude of the current is inferred from its absorption of power from the exciting circuit. As the frequency of the exciting circuit is varied, the power absorption goes through a maximum at the natural frequency of the wavemeter. The Q_0 of the wavemeter circuit is a measure of the sharpness of the maximum. If L and C are known for the wavemeter, the exciting frequency for maximum response is determined. The wave-

meter itself may contain a thermal ammeter for determining the current maximum though as this increases R it results in a decrease in Q_0. If the capacity is a variable condenser, the wavemeter can be used to measure frequencies over a certain range. A scale on the condenser can be graduated directly in terms of frequency; it will be a linear scale if the condenser plates are of the straight-line-frequency type (Sec. 1.5). If the inductance and capacity are known, the scale can be immediately calibrated. This is not in general feasible and the scale is calibrated by setting for resonance at a series of known frequencies throughout the range. The series of harmonics generated by a piezoelectric crystal with its fundamental well below the range of the wavemeter is particularly convenient for this purpose.

The capacity of a circuit is generally the simplest parameter to vary in a known way and the variation of the circulating current with capacity is a standard method of determining the other circuit parameters. Let i_0 be the effective current at resonance and i the effective current for some arbitrary value of the reactance X. Then

$$i_0^2 = \frac{V^2}{R^2} \qquad \text{and} \qquad i^2 = \frac{V^2}{R^2 + X^2}$$

hence

$$\frac{i_0^2 - i^2}{i_0^2} = \frac{X^2}{R^2} = Q^2\left(1 - \frac{\omega_0^2 C_r}{\omega^2 C}\right)^2$$

Here C_r is the resonance setting of the condenser and the potential applied to the circuit is assumed to remain constant. If the exciting frequency is also constant, $\omega = \omega_0$, and $Q = Q_0$ and writing Q_0 explicitly

$$Q_0 = \pm \frac{C}{C - C_r}\sqrt{\frac{i_0^2 - i^2}{i^2}}$$

Hence a measurement of i_0, C_r, i, and C determines the value of Q for the circuit at the resonant frequency. For large values of Q, C_r is very critical and difficult to determine; hence it is more satisfactory to find the two values of the capacity, one greater and one less than C_r for which the current has the same value i. Calling these two values C_1 and C_2, of which it will be assumed that C_1 is the greater, and eliminating C_r from the expressions for Q_0 in terms of these capacities

$$Q_0 = \frac{C_1 + C_2}{C_1 - C_2}\sqrt{\frac{i_0^2 - i^2}{i^2}} \tag{13.9}$$

If i is chosen to be $i_0/\sqrt{2}$, the radical reduces to unity and the expression is a particularly convenient one for the determination of Q_0. It is evident that it is not necessary to know the absolute values of the currents

or capacities but only their ratios to determine Q_0. If ω_0 and C_r are known with sufficient accuracy, the relation $\omega_0^2 = 1/LC_r$ may be used to determine L and hence R may be found from $Q_0 = \omega_0 L/R$. For an accurate determination of either R or L this method suffers from the difficulty in the accurate determination of C_r. On taking the quotient of the expression for Q_0 for the capacities C_1 and C_2, C_r is given with greater accuracy by $C_r = 2C_1C_2/(C_1 + C_2)$. The exciting angular

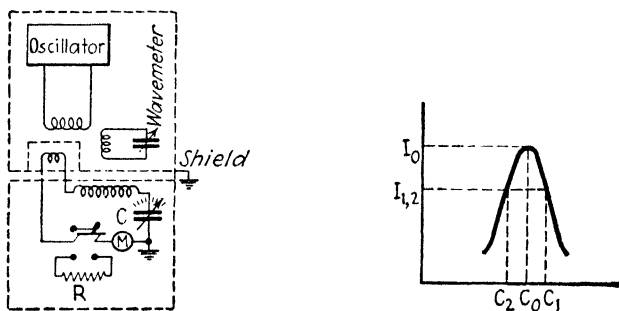



FIG. 13.5.—Measurement of resistance by the capacity-variation method.

frequency ω_0 can generally be determined with great accuracy and the expressions for the inductance and resistance become

$$L = \frac{1}{2\omega_0^2}\left(\frac{1}{C_2} + \frac{1}{C_1}\right) \tag{13.10}$$

and

$$R = \frac{1}{2\omega_0}\left(\frac{1}{C_2} - \frac{1}{C_1}\right)\sqrt{\frac{i^2}{i_0^2 - i^2}} \tag{13.11}$$

Measurements of the parameters, particularly at high radio frequencies, present technical difficulties. Stray capacities are the principal source of error and they must be minimized by adequate shielding. Figure 13.5 illustrates a typical circuit for measuring a radio-frequency resistance. An oscillator coil induces an emf. in a small pickup coil in series with the measuring circuit. The frequency is determined with a wavemeter also loosely coupled to the oscillator. The mutual reactance between these circuits should be very small so that the power dissipated in the wavemeter and measuring circuit is negligible in comparison with that which the oscillator can supply without materially affecting its output or frequency. A shield of parallel wires connected together at one end only (like a comb) will absorb a negligible amount of energy and reduce the capacitative interaction between the oscillator and pickup coil. This equipment is placed in a grounded metal case to shield it from the rest of the apparatus and prevent the induction of emfs. in the measuring circuit except in the pickup coil itself. The leads to the

pickup coil should be symmetrical and leave the shield close together to reduce differential capacitative effects. The measuring circuit itself in general consists of an inductance, capacity, resistance, and a meter. For accurate measurements the meter should be of the thermocouple type. The condenser should be supplied with a grounded shield to reduce the effects of the body capacity of the operator. With the switch in its upper position, L and R of the measuring circuit can be determined by the variation of C, as described in the previous paragraph. If the switch is closed in its lower position, the alteration in the current at resonance can be used to determine the additional resistance introduced. Or if a standard radio-frequency resistance is available, the resistance of the circuit itself may be checked by this measurement. If an unknown impedance is connected across the lower terminals of the switch, its radio-frequency resistance and reactance can be determined from the values of C_1 and C_2 with and without the impedance and the parameters of the measuring circuit itself. If the oscillator frequency is variable, this method can be used to determine the variation in effective resistance and inductance of a circuit element as a function of the frequency.

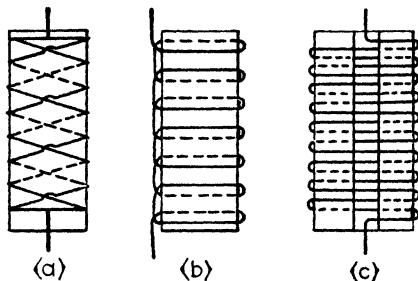



FIG. 13.6.—Typical high-frequency resistances wound to reduce inductance and capacity. (*a*) Ayrton-Perry; (*b*) reversed loops (halfhitches); (*c*) figure eight.

13.3. Parameters of Circuit Elements as Functions of the Frequency.—Even simple circuit elements show a variation in their effective values of R, L, and C as the frequency is changed, and this variation is both of theoretical interest and practical importance. Consider first a straight wire. It has a certain inductance and capacitance per unit length which are determined by its situation and the nature of the rest of the circuit. However, these are generally of negligible importance below frequencies of the order of 10^7 or 10^8 per second. The resistance of the wire is determined primarily by its length, cross section, resistivity, and temperature. However, at high frequencies the distribution of current ceases to be uniform over the cross section and the apparent resistance increases with the frequency. This is due to the tendency for the current to redistribute itself in such a way as to be enclosed by the minimum of flux. This phenomenon is known as the skin effect and has been discussed in Sec. 10.2. A case of great practical interest is that of a long straight wire of circular cross section. The effective high-frequency resistance depends on the parameter $\left(\frac{\mu\sigma\omega\mu_0 a^2}{4}\right) = \alpha$, where μ is the permeability, σ the conductivity, and a the radius of the wire. When α is large, the effective resistance is $\sqrt{\alpha/2}$ times the direct-current resistance. At the more common extreme of small skin effect the percentage change in resistance is $\alpha^2/12$.* Table I gives the maximum wire radius in millimeters

* Problem 23, Chap. X.

for various substances at the frequencies below which the resistance is within 1 per cent of its direct-current value.

From this table it is evident that a large iron wire will exhibit an appreciable skin effect at audio frequencies. The change in resistance of copper wires greater than 0.1 mm. in diameter must be taken into account above a megacycle. On the other hand, carbon filaments, such as those in old incandescent lamps, can be used as resistance standards with a negligible correction well above 100 megacycles. It is, of course, necessary to take into account the variation of resistance with temperature in the case of incandescent lamps. Since the brightness of a lamp is a function of the power consumed, it may be used to measure radio-frequency power or current

TABLE I[1]

Substance	σ (mho per meter)	ν (per second)					
		10^3	10^4	10^5	10^6	10^7	10^8
Carbon	3×10^4	72	23	7.2	2.3	0.72	0.23
Nichrome	10^6	12	3.8	1.2	0.38	0.12	0.04
Constantan	2×10^6	2.7	0.87	0.27	0.09	0.03	0.01
Copper	5.8×10^7	1.6	0.51	0.16	0.05	0.02	
Iron (μ = 100)	10^7	0.4	0.13	0.04	0.01		

[1] Tabular entries are maximum radius in millimeters for which resistance is within 1 per cent of direct-current value for frequency at head of column.

The lamp and a photronic cell are mounted in an enclosed but ventilated container. The deflection of a microammeter in series with the photronic cell is a measure of the power consumed by the lamp. The calibration is obtained at a commercial frequency

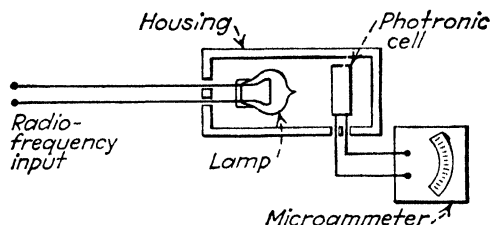



FIG. 13.7.—Use of an incandescent lamp and photronic cell for measuring radio-frequency power or current.

or with direct current. At very high frequencies the lamp should be debased to reduce shunting capacity and the filament should be of the straight type to minimize inductive effects for the accurate measurement of current. Power consumption is, of course, independent of these factors. Thermistors are also used for the measurement of high-frequency power as described in Sec. 5.8.

Well-constructed condensers employing a good quality dielectric show very little change of capacity with frequency. At very high frequencies lead and plate inductances become appreciable, leading to a slight decrease in effective capacity. The losses in a condenser are principally those associated with the dielectric. For large capacities solid or liquid dielectrics are necessary and these show small power losses which increase rapidly with the temperature. The power factor of a condenser or dielectric, which is ωCR, where R is the effective series resistance (Sec 7.6), is so

small that it is entirely negligible for many purposes. Also the losses for most materials vary in such a way with the frequency that the power factor is approximately a constant over very wide range and it or its reciprocal, the Q value, is a significant constant for the particular dielectric. Values for representative substances were given in Table III, Chap. III. A gas is the most satisfactory material for a dielectric at high frequencies. It is used under pressure for condensers that must withstand a high voltage. Solid dielectrics are necessary for mechanical support, but they should be kept out of the intense electric field. Variable air condensers of the rotating-plate type are common for capacities below about 10^{-9} farad. For these it is found that ωRC^2 rather than ωRC is approximately a constant. For good condensers this is of the order of 6×10^{-14}, and hence the losses are extremely small.

Inductances, on the other hand, always have much larger losses associated with them and the effective values of their inductance and resistance are subject to a wide variation with frequency. Consider first low frequencies in the power and audio range. Here it is possible to make use of the high μ value of ferromagnetic materials to construct large inductances. However, it has been seen previously that μ is not independent of the current and this renders all simple calculations rather inaccurate. Also the effective inductance varies with the frequency owing to the magnetic skin effect. It was seen in Sec. 10.2 that the penetration into a conductor of the magnetic induction due to an alternating field decreases with the frequency. This causes unequal flux densities in the core, and in most instances results in a decrease in effective permeability with frequency. The effective permeability may be of the order of several thousand for low frequencies but decreases to the order of 10 at radio frequencies even for carefully prepared dust-core coils. The effective resistance is the power loss divided by the square of the current. For iron-core coils the power loss is made up of three parts: ohmic losses in the winding and hysteresis and eddy current losses in the core. The first is constant, the second varies as the first power, and the third as the second power of the frequency. The result is a wide variation in effective parameters with frequency and a coil is generally designed for a specific frequency range. Over a limited range of frequencies the Q value of an inductance is often approximately constant and for well-designed elements generally lies in the range from 50 to 150. At the higher audio frequencies the interturn capacity frequently becomes of importance and leads to resonance effects.

For frequencies above approximately 10^4 the disadvantages of ordinary iron cores more than outweigh their advantages, and inductances are wound on ordinary insulating materials or are self-supporting. The variation in true inductance with frequency is very small and may generally be neglected but the interturn capacity brings in a capacitative component of the reactance which is not in general calculable. The losses are due to the dielectric and high-frequency resistance. The skin effect in a coil is much greater than for a straight wire owing to the magnetic field of neighboring turns. The result of these factors is that in the radio-frequency range also the Q value is approximately constant and a more significant coil parameter than the resistance. For well-designed coils it generally lies in the range from 100 to 300. In the frequency range below, about 10^6 the losses may be reduced by employing multistrand insulated wire (Litz) which reduces the skin effect by producing a more uniform current distribution. At higher frequencies either solid wire or tubing is preferable.

The design of inductances for the maximum L and minimum R and distributed C is largely an empirical procedure and for further information on the considerations involved reference should be made to radio engineering or communication handbooks. Certain of the inductance formulas in Sec. 9.6 are useful in particular instances. Also a number of approximate formulas have been developed for the more common coil

shapes. In the case of a single-layer solenoid, such as that of Fig. 13.8*a*, the inductance is given by

$$L = \frac{0.0395a^2n^2}{b} K \text{ microhenrys}$$

where n is the total number of turns and a and b are the radius and length, respectively, in centimeters. K is a factor depending on the ratio $2a/b$, which is given by the graph of Fig. 13.9. The inductance of the multilayer coil of Fig. 13.8*b* is given by

$$L = \frac{0.32a^2n^2}{6a + 9b + 10c} \text{ microhenrys}$$

where the dimensions are in centimeters. The accuracy is of the order of 1 per cent if the factors in the denominator are approximately equal. The effective inductance,

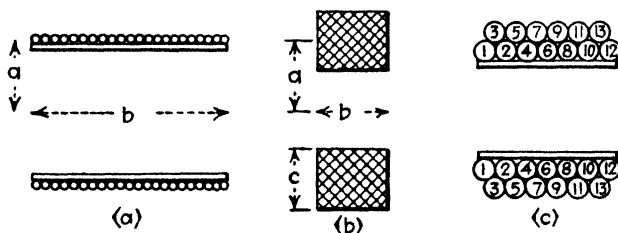



FIG. 13.8.—Familiar forms of radio-frequency inductances.

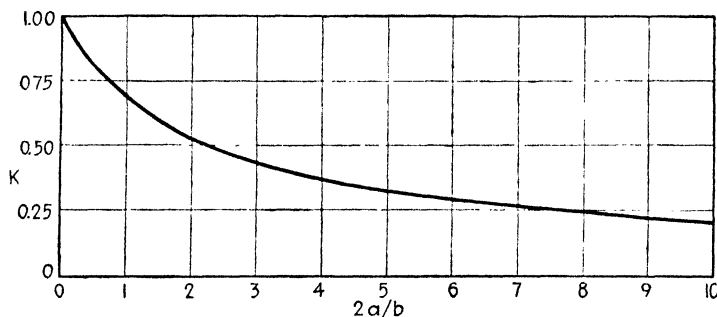



FIG. 13.9.—Factor for determining single-layer solenoid inductance.

which would be obtained, for instance, by the method of the previous section, differs from the calculated inductance because of the distributed capacity of the windings. If a coil is resonated at an angular frequency ω by means of a capacity C

$$\frac{1}{\omega^2} = L(C + C_d) = L_eC$$

where C_d is the distributed capacity of the coil and L_e the effective value of the inductance. On making two measurements, one at ω_1 and one at ω_2, and eliminating C_d the true inductance is seen to be

$$L = \frac{1}{C_1 - C_2}\left(\frac{1}{\omega_1^2} - \frac{1}{\omega_2^2}\right)$$

C_d can also be determined from these measurements. For a closely wound single-layer solenoid the capacity in micromicrofarads is of the order of the diameter of the

coil in centimeters. For multilayer coils the distributed capacity becomes quite large, though it can be kept to a minimum by the bank winding illustrated in Fig. 13.8c. The natural frequency of a coil which is determined by the product LC_d generally lies in the ordinary radio-frequency range. Above this frequency the coil behaves as a capacity rather than an inductance, and long before this frequency is reached the current ceases to be uniform throughout the coil. The current and potential distributions resemble those of the velocity and amplitude of oscillation of air in a sounding organ pipe. The fundamental mode of oscillation of a single-layer solenoid with free ends in which an oscillating emf. is induced is that in which the ends are current nodes and potential antinodes and the center is a potential node and a current antinode. In common with all distributed parameter systems such a solenoid is multiply periodic and will resonate at harmonic frequencies which are approximately integral multiples of the fundamental. If one end of the coil is grounded, this point becomes a current antinode at resonance. This is the most common form of the so-called *Tesla coil* for the demonstration of high-frequency phenomena. It is analogous to the organ pipe with one closed end.[1]

13.4. Special Forms of Simple Series-parallel Circuits.—The linear relation between **i** and **V** [Eq. (13.3)] permits series and parallel combinations of circuit elements to be handled very simply in the complex form. The complex impedance of elements in series is the sum of their separate complex impedances. Similarly the impedance of elements in parallel is the reciprocal of the sum of the reciprocal impedances. The situation is exactly analogous to that of direct-current resistances. Typical series circuits have been considered in a previous section and here a few of the more useful parallel and series-parallel types will be discussed. If $\mathbf{z}_1$ and $\mathbf{z}_2$ are the impedances of two branches, their impedance in parallel is

$$\mathbf{z}' = \frac{\mathbf{z}_1\mathbf{z}_2}{\mathbf{z}_1 + \mathbf{z}_2}$$

In terms of the resistance and reactance of each branch

$$\mathbf{z}' = \frac{(R_1 + jX_1)(R_2 + jX_2)}{(R_1 + R_2) + j(X_1 + X_2)} \tag{13.12}$$

Separating into real and imaginary parts

$$\mathbf{z}' = R' + jX'$$

where

$$R' = \frac{(R_1 + R_2)(R_1R_2 - X_1X_2) + (X_1 + X_2)(X_1R_2 + X_2R_1)}{(R_1 + R_2)^2 + (X_1 + X_2)^2} \tag{13.13}$$

$$X' = \frac{(R_1 + R_2)(X_1R_2 + X_2R_1) - (X_1 + X_2)(R_1R_2 - X_1X_2)}{(R_1 + R_2)^2 + (X_1 + X_2)^2} \tag{13.14}$$

[1] For a further discussion of high-frequency measurements and the variation of circuit parameters see Moullin, "Radio-frequency Measurements," Charles Griffin & Co., Ltd., London, 1931; Hund, "High-frequency Measurements," McGraw-Hill Book Company, Inc., New York, 1933.

These are the expressions for the resistance and reactance for two *L-R-C* circuits in parallel as shown in Fig. 13.10. The condition of resonance is defined as that for which X' vanishes. Equating X' to zero, Eq. (13.14) yields the resonant condition

$$X_1X_2(X_1 + X_2) = -(X_1R_2^2 + X_2R_1^2) \tag{13.15}$$

In general this is a cubic equation in ω^2 and the three roots give the three resonant frequencies. In the limit of $R_1 = R_2 = 0$, R also vanishes and the impedance is a pure reactance given by $X_1X_2/(X_1 + X_2)$. In this case the resonant frequencies are given by Eq. (13.15) as those determined by $X_1 = 0$, $X_2 = 0$, or $X_1 = -X_2$. The first two are the resonant frequencies of the circuits separately and result in a zero impedance; the third evidently produces an infinite impedance. In any actual circuit the resistances cannot be neglected in the neighborhood of resonance, but these three tendencies are observable if the circuits are periodic. In general extreme values of the impedance Z' do not occur at resonant frequencies in distinction to the simple series-resonance situation. The general case will not be considered further, but our attention will be confined to special cases of particular interest.

FIG. 13.10.—Parallel impedances.

In the first place, if the resistances in the two branches are equal, Eq. (13.15) simplifies to $X_1 = -X_2$ and $X_1X_2 = -R^2$. The first

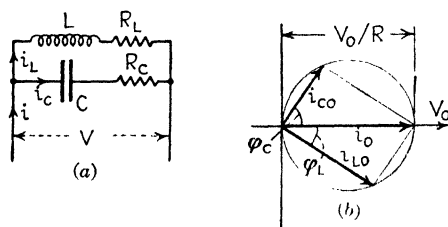



FIG. 13.11.—(*a*) Parallel circuit in which both branches contain resistance. (*b*) Current vectors as a function of the applied frequency when $R_L = R_C = R = \sqrt{L/C}$.

of these equations represents resonance at the natural frequency of the series circuit. The second condition is of particular interest if one reactance is purely inductive and the other purely capacitative, as indicated in Fig. 13.11. In this case the product X_1X_2 is $-\dfrac{L}{C}$ and is independent of the frequency. Thus, if the resistances in the two branches are both equal to $(L/C)^{1/2}$, the phenomenon of resonance is absent. The current is a constant independent of the frequency and is in phase with the applied potential. Its maximum value is given by

Eq. (13.13) or the vector diagram of Fig. 13.11 as V_0/R. The most commonly encountered situation is that in which the resistance in the capacitive branch is negligible. In this case Eqs. (13.13) and (13.14) become

$$R' = \frac{R}{\omega^2C^2R^2 + (\omega^2LC - 1)^2}$$

$$X' = \frac{\omega[L - C(R^2 + \omega^2L^2)]}{\omega^2C^2R^2 + (\omega^2LC - 1)^2}$$

From the numerator of X' resonance is seen to occur at the angular frequency

$$\omega = \left(\frac{1}{LC} - \frac{R^2}{L^2}\right)^{\frac{1}{2}} = \omega_0\left(1 - \frac{CR^2}{L}\right)^{\frac{1}{2}} = \omega_0\left(1 - \frac{1}{Q_0^2}\right)^{\frac{1}{2}}$$

where ω_0 is written for $1/\sqrt{LC}$. Thus, if Q_0 is large, resonance occurs very nearly at $\omega = \omega_0$. On substituting this value of ω in the expression

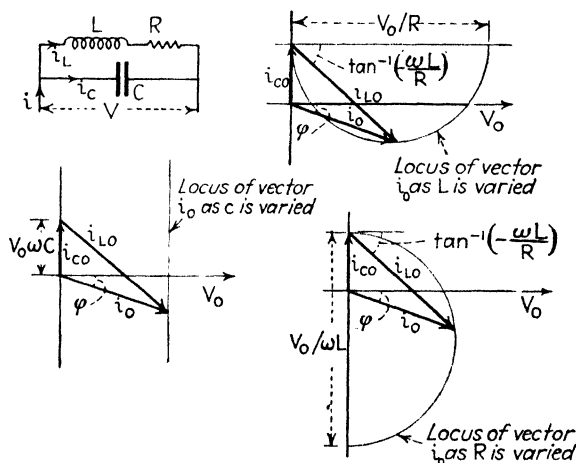



FIG. 13.12.—Current vectors in a parallel circuit as functions of the circuit parameters.

for R', it is seen that the effective resistance at resonance is given by $R' = L/RC$. The effective resistance at resonance is therefore inversely proportional to the resistance of the inductance. This is not in general equal to the maximum impedance Z' of the circuit, but the two conditions are closely the same if the value of Q_0 is large. In terms of Q_0, R' at resonance is RQ_0^2. The vector analysis of the circuit is shown in Fig. 13.12. The total current i to the circuit is the sum of i_c and i_L through the two branches. i_c is always along the imaginary axis and the circles indicate the way in which the total current varies with L, R, and C. It is evident from the diagram for the variation with C that the maximum current occurs at resonance. Figure 13.13 shows the nature of the varia-

tion of the current, which is inversely proportional to Z', with frequency for three representative values of Q_0. The curves are not symmetrical on either side of the resonant point, and for $Q_0 = 1$ it is evident that the minimum current occurs at a considerably lower angular frequency than ω_0. The analytic expression for Z' is

$$Z' = (R'^2 + X'^2)^{1/2} = \left(\frac{R^2 + \omega^2 L^2}{\omega^2 C^2 R^2 + (\omega^2 LC - 1)^2}\right)^{1/2} \tag{13.16}$$

From this expression the values of the parameters that render Z' a maximum can be determined by differentiation.

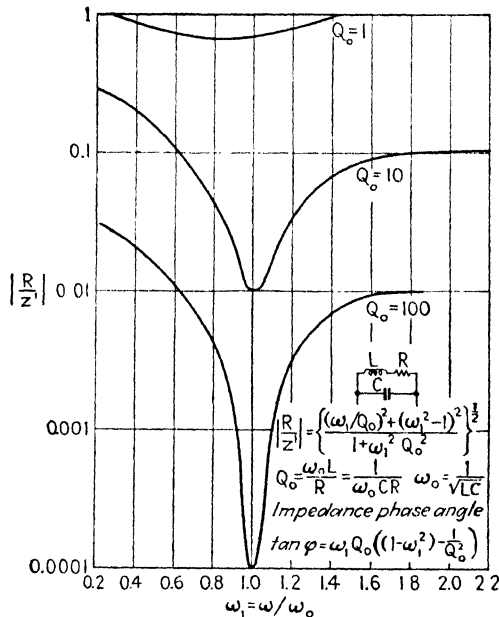



FIG. 13.13.—Variation of the current to a parallel resonant circuit as a function of the frequency.

The very large effective resistance that is presented by a high Q parallel circuit at resonance finds many important applications. Its discrimination or differential response with frequency is of the same order as the series circuit and in applications they play complementary roles. The series circuit presents a very low resistance path for the resonant frequency and the parallel circuit presents a very high one. The high resonant resistance of the parallel circuit makes it a particularly suitable load in vacuum-tube circuits where the plate resistance of the tube is generally high. A form frequently encountered at very high frequencies is illustrated in Fig. 13.14. This is really a circuit of inductance in parallel with an inductance and capacity with mutual inductance between the two branches. Since the separate inductances and the

mutual inductance are not generally calculable the adjustment of the circuit is empirical. However, an elementary analysis shows that the circuit retains the characteristics of that of Fig. 13.12 and presents a high resistive impedance at the angular frequency $1/\sqrt{LC}$. As the tap d is moved along the coil, the value of this effective resistance varies from its maximum value of L/RC to zero. The variation is not linear, but the adjustment is a very convenient one for obtaining any particular value of the effective resistance.

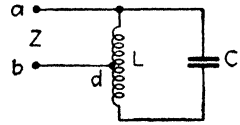



FIG. 13.14.—Tapped inductance.

A useful circuit for the measurement of a large inductance is shown in Fig. 13.15. If the inductance and capacity are such as to satisfy the condition $\omega^2 = 1/2LC$, Z' of Eq. (13.16) reduces to $Z' = (2L/C)^{1/2} = 1/\omega C$. This is also the value of Z' if the inductive branch is not present. Therefore, if opening or closing the key K does not alter the effective current i, this condition between ω, L, and C is fulfilled. A large capacity which is variable in sufficiently small steps is necessary for the measurement. If the inductance has an iron core, an ammeter should be included in

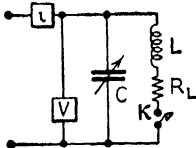



FIG. 13.15.—Measurement of a large inductance.

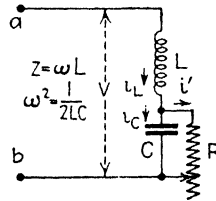



FIG. 13.16.—Phase-shifting circuit (impedance of constant magnitude and variable phase).

that branch to determine the effective current for which the inductance has the value determined. Any pair of the quantities ω, C, and Z' may be known, for when i is independent of the key K

$$L = \frac{1}{2C\omega^2} = \frac{Z'}{2\omega} = \frac{CZ'^2}{2}$$

Another circuit that has interesting properties at particular frequencies is that of Fig. 13.16. The circuit exhibits the same properties (neglecting the resistance of the inductance) if the resistance R is across L instead of C. The complex impedance presented by the terminals a-b is

$$\mathbf{z} = j\omega L + \frac{R}{1 + j\omega CR}$$

which after some reduction can be written

$$\mathbf{z} = Ze^{j\varphi}$$

where

$$Z = \left[\frac{R^2(1 - 2\omega^2LC)}{1 + \omega^2C^2R^2} + \omega^2L^2\right]^{1/2}$$

and

$$\varphi = \tan^{-1}\frac{\omega L}{R}\left[1 - \frac{C}{R}R^2(1 - \omega^2LC)\right]$$

From the expression for Z it is evident that if $\omega^2 = 1/2LC$, the absolute magnitude of the impedance presented by the circuit is equal to ωL and is independent of the setting of R. Hence the effective current to the circuit is $V_e/\omega L$ and does not depend on R. On the other hand, when R is varied from 0 to ∞ φ, which is $\left(\frac{\pi}{2} - 2\tan^{-1}\omega RC\right)$, changes continuously from $\frac{\pi}{2}$ to $-\frac{\pi}{2}$. Thus the arrangement is very useful for changing the phase of a circuit vector without affecting its magnitude. The potential appearing across L is equal in magnitude to V_0 and leads i by $\pi/2$. This circuit has even more interesting properties at the frequency corresponding to $\omega^2 = 1/LC$. From the previous analysis it may be seen that at this frequency the current through R is independent of the resistance and is determined only by the values of V_o, L, and C. This may also be shown by simply considering the current i' through R for this frequency. Since $X_L = -X_c$, this may be written

$$\mathbf{i}' = \mathbf{i}_L - \mathbf{i}_c = \frac{\mathbf{V}_L}{jX_L} - \frac{\mathbf{V}_c}{jX_c} = -\frac{j(\mathbf{V}_L + \mathbf{V}_c)}{X_L} = \frac{j\mathbf{V}}{X_c}$$

Thus the current in its complex form is $\mathbf{i}' = \mathbf{V}\sqrt{C/L}\,\epsilon^{-\frac{j\pi}{2}}$; its magnitude is independent of R and it lags the potential applied to the circuit by $\pi/2$. This is an example of the so-called *constant-current* type of circuit. Its analysis algebraically or graphically in terms of complex vectors shows how i_L, i_c, V_L, and V_c vary as functions of R. The power dissipated in R being equal to i_e^2R is directly proportional to R rather than inversely proportional to R $(V_e/R)^2$ as it is in the more familiar circuits that operate at a constant potential. This gives the circuit many important features and it is particularly valuable for operating gas discharges.[1] The

[1] It was seen in Sec. 5.6 that the stable operation of a device with a falling characteristic in a circuit of constant potential requires a large series resistance. In the notation of Eq. (5.10) the external resistance must be greater in magnitude than (dV/di), the negative slope of the characteristic, in order that the over-all dynamic resistance should be positive. This leads to unnecessary losses. An arc or gas-discharge device will operate stably in a constant-current circuit with no series resistance. (See Sec. 15.6.) A protective parallel resistance, however, should be provided to prevent the potential across it from rising too high if the arc should cease to conduct.

inverse relation to the constant-potential circuit is brought out by the consideration that the terminals of a constant-current circuit can be shorted with no ill effect, but if the circuit opens, the potential difference across the terminals rises and endangers the other circuit elements.

13.5. Bridge and Balanced Circuits.—The Wheatstone bridge, as a circuit for the comparison of resistances, was discussed in Sec. 4.6. Applications in nonlinear circuits arose in Secs. 5.4 and 5.9, and it formed the basis of the push-pull and balanced vacuum-tube circuits of Sec. 7.4. Its uses in general alternating-current circuits are even more important and warrant its separate consideration. One application is in the comparison of impedances and another is in balancing networks which are designed to isolate one branch of a circuit from the influence of emfs. generated in another. The typical impedance bridge is illustrated in Fig. 13.17. It is closely analogous to the Wheatstone bridge for the comparison of resistances. The arms, however, are composed of impedances instead of resistances and the battery and galvanometer are replaced by suitable alternating-current devices. The bridge may be used at either power, audio, or radio frequencies. At commercial frequencies a small transformer is a suitable power supply. A thermocouple or rectifier and direct-current galvanometer may be used as the zero current or balance indicator, but a more sensitive device is the vibration galvanometer. This is an instrument of the D'Arsonval type, but with a second fiber attached to the coil from beneath. Both fibers are heavier than in the ordinary instrument and the coil is held under considerable tension between them. The natural period of the coil is adjusted by varying this tension until it is in resonance at the frequency which is applied to the bridge. In this condition a light spot reflected from the mirror acquires a large amplitude of vibration for a very small potential applied to the instrument terminals. Balance is indicated by quiescence. At audio frequencies an audio oscillator or electrically driven tuning fork is a suitable power source and a rectifier and direct-current meter or headphones with or without amplification can be used for determining the balance point. At higher frequencies any of the standard radio-frequency oscillators can be used to supply the power and one or more stages of amplification followed by a rectifier and direct-current meter provide the most satisfactory balance indicator. If the bridge is one

FIG. 13.17—Schematic impedance bridge (Wagner ground in dashed lines).

whose balance depends on frequency, it may be necessary to insert filters in either the power supply or balance indicator to eliminate harmonics.

At commercial frequencies few precautions are necessary and the circuit may be limited to the solid lines of Fig. 13.17. It is evident that the balance condition at which the detector junctions are at the same potential is given by

$$\frac{\mathbf{z}_1}{\mathbf{z}_2} = \frac{\mathbf{z}_3}{\mathbf{z}_4} \qquad \text{or} \qquad \mathbf{z}_1\mathbf{z}_4 = \mathbf{z}_2\mathbf{z}_3 \tag{13.17}$$

Since the real and imaginary terms must be equal separately to ensure equality of both phase and amplitude, this represents the two equations

$$R_1R_4 - X_1X_4 = R_2R_3 - X_2X_3$$

and

$$R_1X_4 + R_4X_1 = R_2X_3 + R_3X_2 \tag{13.18}$$

The capacity bridge which was mentioned in an earlier section (Fig. 7.23) is a simple and useful form. As a bridge with approximately equal-ratio arms is preferable, it is an advantage to have both a reactance and resistance variable. This is comparatively easy to do in the case of capacities, but inductances that are variable over a considerable range are more difficult to construct. As a consequence inductances are frequently measured in terms of variable capacities. A bridge for that purpose is one due to Owen, in which $\mathbf{z}_1$ is a resistance, $\mathbf{z}_2$ a capacitance, $\mathbf{z}_3$ a resistance in series with the unknown inductance, and $\mathbf{z}\cdot$ a resistance and capacity in series. In this case the parameters are

$$\begin{array}{llll} R_1 = R_1 & R_2 = 0 & R_3 = R_3' + R_L & R_4 = R_4 \\ X_1 = 0 & X_2 = -\dfrac{1}{\omega C_2} & X_3 = \omega L & X_4 = -\dfrac{1}{\omega C_4} \end{array}$$

Here R_L is the resistance of the inductance and R_3' the series resistance in the arm. L and R_L are then given by Eqs. (13.18) as

$$L = R_1R_4C_2 \qquad \text{and} \qquad R_L = \frac{R_1C_2}{C_4} - R_3'$$

One arrangement is to have R_4 and C_4 adjustable, in which case L is proportional to R_1 and the total resistance in arm 3 is inversely proportional to C_4. It is generally more convenient to have C_4 fixed or variable only in steps in which case R_3' and R_4 are generally the variables. This is a very convenient and widely used bridge for inductance measurements.

Bridge measurements of small L's or C's, which require the use of high frequencies to bring the reactances into a convenient range, require

special precautions to avoid the errors introduced by stray capacities between the bridge arms and between the arms and ground. Electrostatic shields between the transformer windings are indicated in Fig. 13.17 and the dashed circuit in that figure represents one of the simplest methods of balancing stray ground capacities. The impedances $\mathbf{z}_1'$ and $\mathbf{z}_3'$ with their center point grounded constitute a *Wagner ground.* With the switch in position *a* the bridge is brought to approximate balance. Then the switch is thrown to position *b* and a second balance achieved by varying one of the primed impedances. A little consideration shows that this ensures a balance to ground of the oscillator terminals at that particular bridge setting. On returning the switch to position *a* a much better bridge balance can generally be obtained and a higher

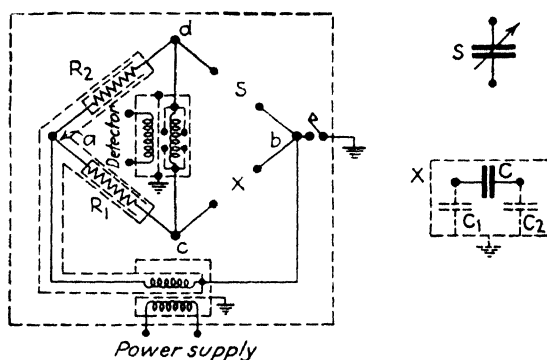



FIG. 13.18.—Equal-ratio-arm completely shielded bridge.

accuracy achieved. This has the disadvantages of frequently requiring several successive approximations to balance and of somewhat restricting the type of impedance measurement that can be made. However, it is a very simple device, and if the bridge is fairly symmetrical, and the reactances are of the same sign, $\mathbf{z}_1'$ and $\mathbf{z}_3'$ can be the two portions of a simple resistance potentiometer.

A more satisfactory but more elaborate alternative to the Wagner ground is complete electrostatic shielding of the bridge. A shielded equal-ratio-arm bridge is shown in Fig. 13.18. The equal-ratio-arm type has the advantages that a simple reversal of *s* and *x* serves to check the measurement and the symmetry facilitates balances to ground. The entire circuit is in a grounded case and the capacity of the point *a* to ground is eliminated by a separate shielding system. The point *b* may be grounded or not, depending on the measurement that is to be undertaken. The bridge may be used for either capacities or inductances, though a variable inductance is necessary for the latter type of measurement. For measuring the leakage conductance of capacities, *x* and *s* are shunted by large resistances as in the circuit of Fig. 7.23.

Actual capacities are generally not simple, but the terminals of the device have capacities to an enclosing shield, as illustrated schematically by C_1 and C_2 of the unknown x. To measure the actual capacity between the terminals $\left[C + \frac{C_1C_2}{(C_1 + C_2)}\right]$ the points b and c must be balanced to ground. To keep the bridge in balance, c and d must have the same capacity to ground and since a has none, these must each be equal to half the capacity of b to ground. If these capacities are provided by small auxiliary condensers, the bridge balance determines $C + \frac{C_1C_2}{(C_1 + C_2)}$. The capacity between the terminals of x when one is connected to the case ($C + C_1$ or $C + C_2$) can be obtained by connecting the case terminal to the point b and grounding it by means of the switch (auxiliary capacities to c and d are, of course, removed). All three measurements will, of course, determine the capacities C, C_1, and C_2 separately. Various other methods will suggest themselves for making more direct measurements of these quantities. One is to ground b and connect the case (or all terminals between which the capacity is not to be measured) to this point. The terminals of the desired capacity are then connected across bc and the bridge balanced. The c terminal is then moved to d and the bridge rebalanced. The difference in the setting of s for the two balances is then twice the desired capacity.

FIG. 13.19.—Mutual-inductance bridge.

If the terminals of a pair of coils are separately available, the mutual inductance between them can be determined by making one measurement of their self-inductance in series and a second measurement after reversing the terminals of one of the coils (a series-aiding and a series-opposing measurement). The difference between the effective inductances is four times the mutual inductance. However, the method is frequently inapplicable and subject to considerable inaccuracy if the mutual inductance is small in comparison with the self-inductances. An alternative method is that provided by Heydweiller's network which is shown in Fig. 13.19. It is a modified bridge circuit, the balance condition being given by

$$\mathbf{z}_1 i_1 = \mathbf{z}_2 i_2 \qquad \text{and} \qquad \mathbf{z}_3 i_2 = \mathbf{z}_m(i_1 + i_2)$$

where $\mathbf{z}_m$ is the complex mutual impedance ($j\omega M$) between the inductances L' and L''. Eliminating the currents, the condition becomes

$$\mathbf{z}_1\mathbf{z}_3 = -\mathbf{z}_m(\mathbf{z}_1 + \mathbf{z}_2)$$

or in terms of the circuit parameters

$$M = -R_1R_3C_2 \quad \text{and} \quad L' = (R_1 + R_2)R_3C_2$$

where R_3 is the total resistance in that branch. If R_1 and R_3' are variable, the adjustment is generally most convenient. Both M and L' are determined in terms of the other parameters. R' is determined by a separate measurement. A reversal of the coils checks the measurement of M and also determines L''. The relative sense of winding of L' and L'' is important; the negative M indicates series opposing.[1]

Another application of the bridge or tetrahedral circuit is in the practical construction of constant-current circuits. The so-called *monocyclic square* is shown at the right in Fig. 13.20. The simple analysis neglecting the resistance of L is the same as that in connection with Fig. 13.16 and i' is equal to $\sqrt{C/L}\,V$ and is $\pi/2$ out of phase with V. The product CL is determined by the frequency at which the circuit is to be used. Thus at 60 cycles LC must be equal to 7 in henrys and microfarads. To this approximation i' is entirely independent of R, and since no power is consumed in the square, $iV = i'V'$ and $i = \sqrt{C/L}V'$. In actual practice losses in the inductances cannot be neglected, and since the inductance of an iron-core coil is a function of the current through it, the resonant condition is strictly fulfilled for only one value of R. However, if the inductors contain large air gaps, the detuning with current is not serious. If R_L and R are both small in comparison with the reactances, the current i' is given approximately by

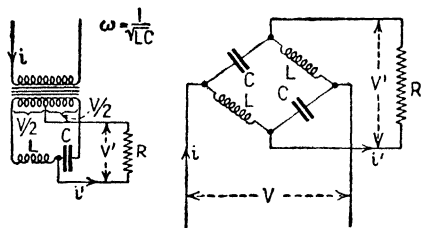



FIG. 13.20.—Representative constant-current circuits.

$$i' = V\sqrt{\frac{C}{L}}\left\{1 - \frac{R_L^2C}{2L}\left[\left(1 + \sqrt{\frac{C}{L}}\frac{X}{2}\right)^2 - \frac{R}{R_L}\right]\right\} \tag{13.19}$$

where R and X are the resistance and reactance of the load, respectively. An alternative circuit is shown at the left in Fig. 13.20. Half of the square is here replaced by a center-tapped transformer. Except for loss calculations, the characteristics of the circuit are the same as those of the monocyclic square. Applications of the principle to two- and three-phase circuits will immediately suggest themselves.

One of the most important uses of the bridge circuit for balancing or neutralization is in connection with vacuum-tube amplifiers. Since the alternating-current power output of such a device is greater than the

[1] References for bridge measurements: HAGUE, "Alternating Current Bridge Methods," Sir Isaac Pitman & Sons, Ltd., London, 1930; CAMPBELL and CHILDS, "Measurements of Inductance, Capacity, and Frequency," D. Van Nostrand Company, Inc., New York, 1935.

alternating-current power input, any transfer of output power to the input may result in instability. This question will be considered further in connection with oscillating circuits, but for the present discussion it is evident that a minimum of interaction between output and input is desirable. The screen-grid tube (Sec. 7.3) is one device for accomplishing this purpose, but if triodes must be used, the bridge circuit provides an alternative. Two representative simple circuits are shown in Fig. 13.21. The lower diagrams represent capacitance neutralization, the one to the right being the simplified capacitance circuit that is present even when

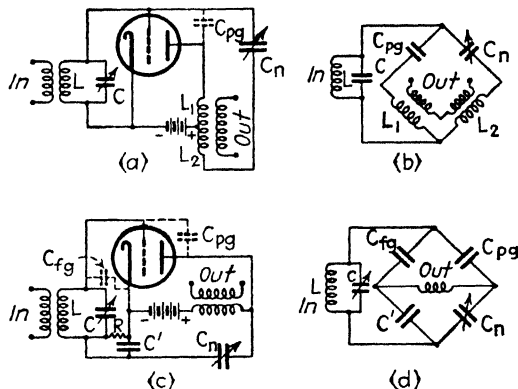



FIG. 13.21.—Use of the balanced circuit to prevent interaction of output and input for triode amplifiers. (a) Neutrodyne circuit; (b) equivalent bridge circuit. (c) Capacitance neutralization; (d) equivalent bridge circuit.

the cathode of the tube is not heated. This is evidently a simple capacity bridge. The adjustment is made by supplying power to the input with the cathode unheated and adjusting the neutralizing condenser C_n till there is no emf. across the output. From the reversibility of the circuit this implies that no emf. introduced across the output terminals will appear across the input. Thus the desired condition is achieved. Heating the cathode has no effect on this balance since it introduces only unidirectional conduction. Batteries which may be necessary to supply the grid potential are not shown; the necessary direct-current path to the grid is, however, indicated by R. Alternatively C' may be a choke coil or inductance operated above its resonant frequency so that it exhibits a capacitative reactance. The upper diagrams represent the neutrodyne circuit which is essentially an inductance-capacity bridge. The considerations are much the same as for the previous case and the adjustment is performed in the same way.

The three-winding transformer was mentioned in connection with constant-current circuits, and it has equally important applications as a component of a bridge circuit in communication work. It is sometimes known as a *hybrid coil* and a general schematic circuit is shown in Fig. 13.22. Consider first that the emfs. V_2, V_3, and V_4 are zero and that

$\mathbf{z}_2 = \mathbf{z}_4$. It is evident from symmetry that an emf. V_1 induces no current in branch 3. The converse must also be the case and an emf. V_3 induces no current in branch 1. Thus power supplied in either branches 1 or 3 is divided equally between branches 2 and 4. If in addition to the condition $\mathbf{z}_4 = \mathbf{z}_2$ the additional restriction $\mathbf{z}_3 = \mathbf{z}_1/n^2$ is met, a simple analysis of the circuit subject to the ideal transformer restrictions shows that an emf. V_2 induces no current in branch 4 and conversely an emf. V_4 induces no current in branch 2. Power originating in either of these branches is divided equally between the other two, namely, 1 and 3. The principal application of this device is in the bidirectional amplifier or telephone repeater illustrated in Fig. 13.23. This arrangement will amplify a signal in either direction, and if the previous restrictions are observed, there will be no interaction between the input and output circuits of the amplifier and hence no tendency for instability. A signal entering from the north (circuit 2) is divided equally between

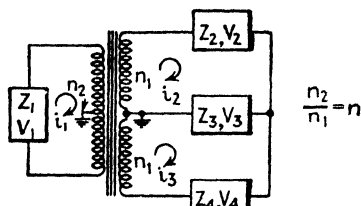



FIG. 13.22.—Form of balanced circuit employing a three-winding transformer or hybrid coil.

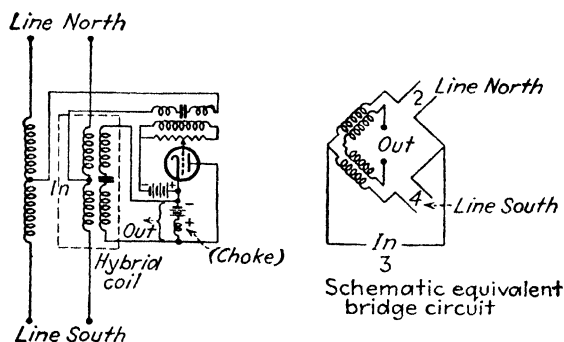



FIG. 13.23.—Bidirectional amplifier or repeater.

the input and output, the latter being lost. That reaching the input (say, circuit 3) appears after amplification at the output (circuit 1) from which it is divided equally between the lines (circuits 2 and 4). Only one-half of the signal power is available for use and one-half of the output goes in the forward direction, but the amplification results in a net over-all gain in signal strength. The condensers are inserted for purposes of direct-current blocking and symmetry. The gain can be controlled by means of the potentiometer across the input of the amplifier. The center-tapped inductance in the left-hand line plays a role only in balancing the lines to ground. The limitation of this simple circuit is that the two line impedances must be equal or the symmetry conditions are not fulfilled. If they are unequal, the circuit may become unstable and

oscillate or "sing." A more complicated circuit involving compensating networks may be used to obviate this requirement.[1]

13.6. General Circuit and Power Considerations.—Before leaving this discussion of special alternating-current circuits, the simple formal extension to the general theory should be pointed out. Provided L, C, and R are constants, the linear relation between **i** and **V** of Eq. (13.3) holds. The fundamental laws of Kirchhoff, of course, apply as well, hence the general linear-circuit theory of Sec. 4.3 can be taken over entirely with the substitution of the admittances $\mathbf{y}_{il}$ for the conductances G_{il} and impedances $\mathbf{z}_{jk}$ for the resistances R_{jk}. The only difference appears in the final reduction of the **y**'s and **z**'s in terms of the frequency and the circuit parameters, which determines the phase as well as the magnitude of the currents in every branch of the network. For the sake of completeness and review of the nomenclature the formal equations will be restated in terms of admittances and impedances.

Shunt Analysis.

$\mathbf{V}_i$ = complex potential of junction i

$\mathcal{E}_{il}$ = complex emf. in the branch between junctions l and i

$\mathbf{y}_{li} = g_{li} + jb_{li}$ = complex admittance of the branch joining junctions l and i

$$\mathbf{y}_{ll} = -\sum_{i \neq l} \mathbf{y}_{li} \qquad \text{and} \qquad \mathbf{I}_l = \sum_{i \neq l} \mathbf{y}_{li} \mathcal{E}_{li}$$

Junction equations:

$$\mathbf{I}_l = \sum_{i=1}^{i=J-1} \mathbf{y}_{il} \mathbf{V}_i \tag{13.20}$$

The solution for the junction potentials are given by the following equations

$$\mathbf{V}_i = \sum_{l=1}^{l=J-1} \frac{\mathbf{B}_{il}}{\mathbf{D}'} \mathbf{I}_l \tag{13.21}$$

where $\mathbf{D}'$ is the symmetrical determinant of the $\mathbf{y}_{il}$'s and $\mathbf{B}_{il}$ is the cofactor of $\mathbf{y}_{li}$ in this determinant.

Series Analysis.

$\mathbf{i}_k$ = complex current assumed to flow in the kth mesh

$\mathcal{E}_j$ = complex emf. encountered in traversing the jth mesh (exclusive of emfs. induced by changing currents in meshes of the network)

$\mathbf{z}_{jk} = r_{jk} + jx_{jk}$ = complex impedance common to meshes j and k

$\mathbf{z}_{jj}$ = total complex impedance in the jth mesh (exclusive of mutual inductances with other meshes of the network)

The quantity **z** with a single subscript is reserved to designate branch impedances.

[1] For a more complete discussion see Johnson, K. S., "Transmission Circuits for Telephonic Communication," D. Van Nostrand Company, Inc., New York, 1931.

Mesh Equations.

$$\mathcal{E}_j = \sum_{k=1}^{k=M} z_{jk} i_k \tag{13.20'}$$

The solution for the mesh currents are given by the following equations:

$$i_k = \sum_{j=1}^{j=M} \frac{\mathbf{A}_{jk}}{\mathbf{D}} \mathcal{E}_j \tag{13.21'}$$

where $\mathbf{D}$ is the symmetrical determinant of the $\mathbf{z}_{jk}$s and $\mathbf{A}_{jk}$ is the cofactor of $\mathbf{z}_{kj}$ in this determinant.

The general circuit theorems of Sec. 4.3, which depend merely on the linearity of the equations and the symmetry of the determinant, of course, apply and frequently aid materially in the general analysis and calculations. The *superposition theorem*, which states that each emf.

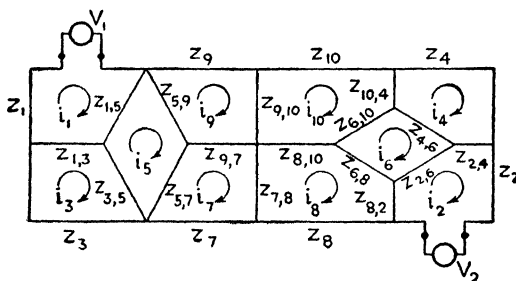



FIG. 13.24.—Typical complex circuit to illustrate series analysis nomenclature and general circuit theorems.

produces currents independently of all the rest, is merely a consequence of the linearity and implies that currents of different frequencies and transients may all be considered separately. The *compensation theorem* states that if any change $\delta\mathbf{z}_k$ is made in an impedance, it will have the same effect on the currents in all other branches as the insertion of an emf. of $-\delta\mathbf{z}_k\mathbf{i}_k$ in series with $\delta\mathbf{z}_k$ (where $\mathbf{i}_k$ was the original current in that branch). This may be seen immediately by making this alteration in the fundamental equations. The *reciprocity theorem* is a consequence of the symmetry of the determinant and is generally stated in a limited form involving the current in one branch and the emf. in another. Writing $\mathbf{V}$ for the emf. instead of $\mathbf{E}$ and assuming all $\mathbf{V}$'s except $\mathbf{V}_1$ are zero, $\mathbf{i}_2 = (\mathbf{A}_{12}/\mathbf{D})\mathbf{V}_1$, where $\mathbf{D}$ is the determinant of the coefficients of the $\mathbf{i}$'s and $\mathbf{A}_{12}$ is the cofactor of $\mathbf{z}_{12}$ in this determinant. If $\mathbf{V}_1$ is transferred to mesh 2, the new $\mathbf{i}_1$ is evidently identical to the former value of $\mathbf{i}_2$ since by symmetry the cofactors $\mathbf{A}_{12}$ and $\mathbf{A}_{21}$ are equal. This is frequently stated in terms of impedanceless ammeters and sources of emf. The fourth useful theorem comes from considering the value of the emf., say $\mathbf{V}'_n$, which when introduced into, say, the nth branch, makes that branch current zero. This is evidently equal to minus the emf. that would occur

across the terminals of the nth branch if it were removed from the circuit. Thus the actual current flowing in the nth branch under the influence of the emf. $\mathbf{V}_n$ may be considered as the current which would be produced by the emf. $\mathbf{V}_n - \mathbf{V}'_n$ applied to a circuit of the impedance viewed from $\mathbf{V}_n$, which is $\mathbf{V}_n/\mathbf{i}_n = \mathbf{D}/\mathbf{A}_{nn}$. This impedance may also be considered as the sum of $\mathbf{z}_n$ and the impedance viewed from the points of attachment of the nth branch to the network. This theorem, it will be recalled, is known as *Thévenin's theorem.*

Though the formal analogy between alternating- and direct-current-circuit analysis is complete, the consideration of power transfer brings out important differences. The current in the circuit of Fig. 13.25 is given by

$$\mathbf{i} = \frac{\mathbf{V}}{\mathbf{z}_1 + \mathbf{z}_2}$$

or

$$i_e = \frac{V_e}{[(R_1 + R_2)^2 + (X_1 + X_2)^2]^{1/2}}$$

This is a maximum as a function of R_1 or R_2 when these quantities are as small as possible. The condition for a maximum current as a function of the reactance is that $X_1 = X_2$, *i.e.*, the resonant condition. The power absorbed by the load impedance $\mathbf{z}_2$ is $R_2 i_e^2$, or

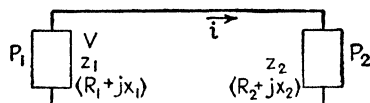



FIG. 13.25.—Considerations affecting power transfer.

$$P_2 = \frac{V_e^2 R_2}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

This is a maximum as a function of R_1 when R_1 is as small as possible and as a function of the reactance when $X_1 = -X_2$. To find the maximum condition as a function of R_2 differentiate with respect to this variable and set equal to zero, which yields

$$R_2 = [R_1^2 + (X_1 + X_2)^2]^{1/2}$$

This is the proper value of R_2 for maximum power transfer. If X_1 or X_2 can also be chosen independently of R_2, the optimum power transfer occurs for $X_1 = -X_2$, in which case the optimum R_2 is given by $R_2 = R_1$. Thus in this particular instance only does the maximum power transfer occur at the equality of resistances which is the direct-current condition for maximum transfer. Otherwise R_2 should always be greater than R_1. P_2 may be expressed as $V_e^2 R_2/Z_t^2$, where Z_t is the absolute magnitude of the total impedance in the circuit. The transfer-power loss is generally expressed in decibels as 10 times the $\log_{10}$ of the ratio of the maximum power that could be transferred subject to the value of R_1 to the actual power transfer. This is evidently

$$\text{Transfer loss} = 10 \log_{10} \frac{P_m}{P} = 20 \log_{10} Z_t - 10 \log_{10} 4R_1R_2 \text{ decibels.}$$

Occasionally for reasons other than those connected with power transfer it is desirable to have the impedances matched, *i.e.*, $R_1 = R_2 = R$ and $X_1 = X_2 = X$. In this case the transfer loss is evidently $10 \log_{10} [1 + (X/R)^2]$. This is seen to increase as X increases.

Under certain circumstances R_2 and X_2 are not independent and the previous analysis does not apply. The most common case is that of an inductance of limited size such as a loud-speaker or some other piece of electrical machinery. Here, since doubling the number of turns of wire implies halving its cross section, both R and L increase at approximately the same rate with the number of turns ($\sim n^2$), and the ratio of X_2 to R_2, or Q, is approximately constant at any particular frequency. Expressing X_2 in terms of the constant Q and differentiating the expression for the power as a function of R_2, the condition for a maximum is seen to be

$$R_1^2 + X_1^2 = R_2^2(1 + Q^2) \qquad \text{or} \qquad |Z_1| = |Z_2|$$

Thus a true match of impedance magnitudes is the desired condition. Of course, if circuit 1 can be chosen, R_1 should be as small as possible and X_1 should equal $-X_2$. Other considerations, such as a minimum variation in power transfer with frequency, are often present to alter these conditions. Finally there is the case in which the impedance of the source and load are fixed and it is desired to insert a transformer which will transfer the maximum power from circuit 1 to circuit 2. Assuming an ideal transformer with n times as many turns on the secondary as on the primary $\mathbf{i}_2 = \mathbf{i}_1/n$ and the effective impedance presented to V is $\mathbf{z}_1 + \dfrac{\mathbf{z}_2}{n^2}$. Hence i_2 is given by

$$i_2 = \frac{i_1}{n} = \frac{Vn}{[(n^2R_1 + R_2)^2 + (n^2X_1 + X_2)^2]^{1/2}}$$

If n is variable, the current in the secondary and hence the power transferred is a maximum when $n^2 = Z_2/Z_1$. This is a useful though very approximate expression for the choice of a transformer for power transfer.

Problems

1. Is a circuit composed of an inductance of 0.5 henry, a capacity of 0.1 μf., with a resistance of 4,000 ohms oscillatory or not? If so, what is the difference between its period and that which it would have if the resistance were negligible?

2. A capacity of 1 μf. is charged to a potential of 100 volts. The battery is then removed and it is shorted by an inductance of 250 millihenrys which has a resistance of 200 ohms. Express the subsequent charge on the condenser as a function of the time. What is the charge on the condenser after 10 complete oscillations? What is the time interval between the shorting of the condenser and the next time that there is no current through the inductance?

3. Show that if a condenser discharges through a critically damped circuit, the current is a maximum $2L/R$ sec. after shorting.

4. A battery of potential V is connected to a circuit made up of an uncharged condenser C and inductance L with resistance R at a time $t = 0$. Show that the initial conditions can be expressed either as $i = 0$, $di/dt = V/L$ at $t = 0$ or $q' = CV$ and $dq'/dt = 0$ at $t = 0$, where q' is the charge on the condenser minus CV. Assuming the circuit to be periodic, find the current in it as a function of the time. Show that the maximum potential difference that appears across the condenser is $V\left(1 + e^{-\frac{\delta}{2}}\right)$, where δ is the logarithmic decrement of the circuit.

5. What is the logarithmic decrement of a circuit made up of 10-μf. condenser and a 5-henry inductance with a resistance of 400 ohms? At what rate would power have to be supplied to this circuit to keep the effective current at 1 amp.? What would then be the maximum potential difference appearing across the condenser?

6. Show that the maximum potential across the condenser of a series-resonance circuit occurs at the frequency $\omega = \omega_0\left(1 - \frac{1}{2Q_0^2}\right)^{\frac{1}{2}}$. Show that the maximum potential across the inductance occurs at $\omega = \omega_0\left(1 - \frac{1}{2Q_0^2}\right)^{-\frac{1}{2}}$.

7. Draw the vector diagrams for the constant-current circuits of Fig. 13.20.

8. Draw the vector diagram for the circuit of Fig. 13.16 both for the angular frequency $1/\sqrt{LC}$ and $1/\sqrt{2LC}$.

9. Show that the effective inductance of two coils of self-inductance L_1 and L_2 and mutual inductance M is $L_1 + L_2 \pm 2M$ if they are connected in series. Show that their effective inductance is $(L_1L_2 - M^2)/(L_1 + L_2 \pm 2M)$ if they are connected in parallel. Examine the cases of $L_1 = L_2$ and $M^2 = L_1L_2$. The latter is called *unity coupling.*

10. Show that the natural frequency of a single-layer solenoid of 100 turns, 4 cm. in diameter and 10 cm. long, is of the order of 7 megacycles.

11. The impedance of a coil is measured at a certain frequency. An ideal condenser is then placed across its terminals and adjusted for resonance at that frequency. If the effective resistance of the combination is n times the coil impedance, show that the Q of the coil is approximately $n\left(1 + \frac{1}{2n^2}\right)$ if n is large.

12. A 10-microhenry inductance is connected in series with a capacity of 100 $\mu\mu$f. Across this combination is an inductance of 12 microhenrys. Neglecting resistances, what are the frequencies of maximum and minimum impedance?

13. Two circuits consisting of an inductance and capacity in series are to be placed in parallel to present a maximum impedance at 500 kilocycles and minima at 490 and 510 kilocycles. Design the circuit neglecting resistances and using values of inductance of the order of 1 millihenry.

14. A 1-henry inductance is connected in parallel with a condenser of 16μf. If the resistance of the inductance is 100 ohms, what resistances must be placed in the two branches to render the effective resistance of the parallel circuit the same for the power frequency and its harmonics? If the line voltage is 125, what current is drawn by the circuit?

15. A capacity C and inductance L with an internal resistance R are connected in parallel. Show that the impedance of the combination is a maximum for the angular frequency

$$\omega = \left(\frac{\sqrt{\frac{2R^2C}{L} + 1}}{LC} - \frac{R^2}{L^2}\right)^{\frac{1}{2}}$$

What is the impedance of the combination at this frequency?

16. Show that the condition for resonance for a parallel circuit of inductance and capacity for variation of the inductance is

$$L = \frac{Q^2}{\omega^2 C(1 + Q^2)}$$

if the resistance of the coil increases with the inductance in such a way that the Q remains constant. Show that the condition for maximum impedance under these conditions is $L = 1/\omega^2 C$. Find the impedance presented by the circuit under these two conditions.

17. Find graphically the locus of the total current vector for the parallel resonant circuit when L is varied at constant Q.

18. Derive Eq. (13.19).

19. A 10-henry inductance is available for the construction of the phase-shifting circuit of Fig. 13.16 on a 60-cycle line. What is the proper value of the associated capacity? If the resistance is variable from 0 to 10,000 ohms, what is the change in phase that can actually be achieved? Plot φ as a function of R at 1,000-ohm intervals.

20. A complex-current wave of the form

$$i = i_0 + i_1 \cos \omega t + i_2 \cos 2\omega t + i_3 \cos 3\omega t$$

is observed to flow to a parallel resonant circuit of a coil and condenser. The coil has a Q of 200, and a capacity of 1,000 $\mu\mu$f. resonates the circuit to the fundamental which is 1 megacycle. What is the trigonometric series for the potential difference appearing across the parallel circuit? What is the consumption of power at the fundamental and at the harmonic frequencies?

21. A vacuum tube has an effective plate resistance of 5,000 ohms. What is the turn ratio of the ideal transformer that should be used to couple it to a loud-speaker with a resistance of 2000 ohms and an inductance of 1 henry? Assume a frequency of 1,000 cycles. What is the transfer loss in decibels (*a*) without the transformer, (*b*) with it?

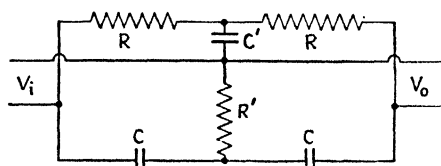
22. What would be the proper value of a series condenser to use in place of the transformer of the previous problem? What would then be the transfer loss at 500 and 2,000 cycles, respectively? How do these compare with the losses without the condenser?

23. A variable frequency can be applied to a high-Q resonant circuit. If the currents at frequencies corresponding to $\omega_0 \pm \delta\omega$ are $1/\sqrt{2}$ times the current at the resonant frequency ω_0, show that the value of Q_0 for the circuit is $\omega_0/2\delta\omega$ approximately.

24. If $\mathbf{E} = E_0^{j\omega t}$, $\mathbf{Z} = Z_0^{j\varphi}$, and $\mathbf{E} = \mathbf{Z}i$, show that the average power being dissipated in $\mathbf{Z}$ [Eq. (13.7)] may be written in the following alternative forms:

$$P = \tfrac{1}{4}(Ei^* + E^*i) = i_e^2 Z_r = \frac{E_e^2}{Z_r}$$

where the asterisk indicates the complex conjugate (change of j to $-j$) and Z_r is the real part of the complex impedance.

25. In the accompanying circuit diagram $R' = R/2$ and $C' = 2C$. If $x = \omega RC$ show that for the open output circuit shown $\dfrac{V_0}{V_i} = \dfrac{x - (1/x)}{4j - (x - 1/x)}$. Plot this ratio as a function of the parameter x and determine the values of x for which the magnitude of the ratio is $1/\sqrt{2}$ times its maximum value (3-db. points). This is known as a *twin-T network* and is very useful as a symmetrical frequency-sensitive coupling circuit.

Prob. 13.25.

CHAPTER XIV

COUPLED CIRCUITS, FILTERS, AND LINES

14.1. Iron-core Transformers at Audio Frequencies.—The transformer, as a device for use at power frequencies, was considered in Sec. 12.6. It is possible to continue to use a ferromagnetic core up through the audio-frequency range, though the losses and the disadvantages due to nonlinearity become of greater importance. Much more stringent requirements are imposed on the properties of the core materials. Various of the available alloys were considered briefly in Sec. 11.8 and it was there pointed out that the accurate reproduction of a wave form requires an alloy with a linear characteristic; otherwise distortion and undesired interaction between separate waves will take place. As these transformers in general operate at very low powers, it is essential to keep the losses to a minimum, which implies a hysteresis loop of small area and a core material of high effective resistance. The small currents which are available mean that the core material must have a high permeability for low magnetizing forces if the transformer is to be an efficient one. The circuits between which these transformers must operate often have a high internal resistance, which largely invalidates the simple transformer assumptions. And finally the large range of frequencies that is to be covered makes the primary inductance, leakage inductance, and distributed capacities of great importance. Their effect is to produce a voltage-transformation ratio quite different from that given by the ratio of the secondary turns to the primary turns.

The capacitative interaction between the windings can be reduced by making the connections in such a way that the contiguous layers of the two windings are at the same alternating potential. If necessary the capacity can be practically eliminated by the interposition of an electrostatic shield between the windings. In consequence this capacitative interaction will be neglected and the two mesh equations applicable to Fig. 14.1 are

$$\begin{aligned} \mathbf{V} &= (\mathbf{z}_1 + j\omega L_p)\mathbf{i}_1 - j\omega M\mathbf{i}_2 \\ 0 &= -j\omega M\mathbf{i}_1 + (\mathbf{z}_2 + j\omega L_s)\mathbf{i}_2 \end{aligned} \qquad (14.1)$$

The quantities of particular interest are the input impedance, $\mathbf{V}/\mathbf{i}_1$ which will be written $\mathbf{z}'$, the secondary current $\mathbf{i}_2$, and the potential across the load, $\mathbf{V}_l$. Eliminating $\mathbf{i}_2$ between Eqs. (14.1) the input impedance is

$$\mathbf{z}' = \frac{\mathbf{V}}{\mathbf{i}_1} = (\mathbf{z}_1 + j\omega L_p) + \frac{\omega^2 M^2}{(\mathbf{z}_2 + j\omega L_s)} \tag{14.2}$$

Eliminating $\mathbf{i}_1$, the secondary current is

$$\mathbf{i}_2 = \frac{j\omega M \mathbf{i}_1}{(\mathbf{z}_2 + j\omega L_s)} = \mathbf{V}\frac{j\omega M}{(\mathbf{z}_1 + j\omega L_p)(\mathbf{z}_2 + j\omega L_s) + M^2\omega^2} \tag{14.3}$$

and the load potential is $\mathbf{i}_l \mathbf{z}_l$, where $\mathbf{i}_l$ may differ from $\mathbf{i}_2$ owing to impedances in the secondary circuit other than those of the load itself. In the ideal power transformer the coil reactances are so large that the series impedances can be neglected in comparison with them. Also, if n is the secondary- to primary-turn ratio, M is proportional to n and L_s to n^2, so $L_p L_s - M^2$ vanishes, leading to

Fig. 14.1.—Schematic transformer with associated impedances.

$$\mathbf{z}' = \mathbf{z}_1 + \frac{\mathbf{z}_2}{n^2} \qquad \text{and} \qquad \mathbf{i}_2 = \frac{\mathbf{i}_1}{n}$$

This is the approximation of Sec. 12.6. In the more general case which is necessary for audio-frequency transformers the square of the mutual inductance is less than the product of the primary and secondary inductances. The *coefficient of coupling*, k, which is always less than unity, is defined by the equation

Fig. 14.2.—Illustration of the effective impedances associated with a transformer.

$$M = k\sqrt{L_p L_s} \tag{14.4}$$

However, for a well-designed transformer k is close to unity and it is profitable to eliminate M from Eq. (14.2) by means of Eq. (14.4) and add and subtract a term $jk\omega L_p$, which permits Eq. (14.2) to be written approximately as

$$\mathbf{z}' = \mathbf{z}_1 + j(1 - k)\omega L_p + \frac{1}{\dfrac{1}{(jk\omega L_p)} + \dfrac{n^2}{k[\mathbf{z}_2 + j(1 - k)\omega L_s]}}$$

if

$$\mathbf{z}_2 >> j(1 - k)\omega L_s$$

This is evidently the impedance of the series-parallel circuit of Fig. 14.2. As $(1 - k)$ is small, the second term, which is part of the leakage reactance, is generally small and the shunt reactance $jk\omega L_p$ is large.

In addition to the leakage reactances there is resistance associated with both coils and also distributed capacity in each. Finally there are certain

core losses which are generally small and which can be represented by a resistance across the primary. Thus the approximate circuit for the transformer can be drawn as in Fig. 14.3. Resistances and reactances to the right of the dashed line are in the secondary, and if they are to be considered as part of the primary circuit, they must be multiplied by L_p/L_s, which is $1/n^2$. When multiplied, they will be designated by a prime in subsequent figures. The potential difference applied to the circuit to the right of the dashed line is n times that appearing across the inductance kL_p. The shunt resistance representing the core losses is generally much larger than the inductive reactance or effective secondary impedance and in the following discussion it will be neglected.

Transformers perform two general types of service at audio frequencies. They are used for impedance-matching purposes to transfer power from one circuit to another and also for potential amplification with

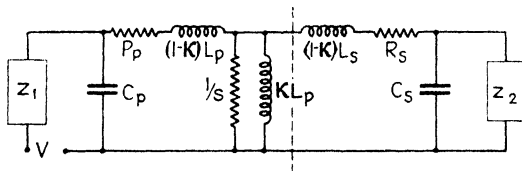



FIG. 14.3.—Approximate schematic circuit of the complete transformer.

the high-resistance input circuit of a vacuum tube in the secondary. In the first application the primary- and secondary-circuit impedances, $\mathbf{z}_1$ and $\mathbf{z}_l$, are generally comparatively small pure resistances and the capacities can be neglected. If the transformer is designed to work in the neighborhood of a certain angular frequency ω'', the circuit in this frequency region, for power purposes, may be thought of as simply the resistances R_1, R_p, R'_s, and R'_l in series with the applied potential V. For the power consumed in the secondary circuit is $(R_s + R_l)i_2^2$; or since $i_2 = i_1/n$, this is the same as the primed resistances multiplied by the square of the primary current. The transformer efficiency is the ratio of the power delivered to the load to the total power delivered to the transformer, or

$$\text{Efficiency} = \frac{R'_l}{R_p + R'_s + R'_l}$$

By suitable transformer and circuit design this can be made as large as 90 per cent. Since a well-designed transformer has R_p approximately equal to R'_s, this implies that the circuit should be designed so that the load resistance is about 20 times the effective resistance of the transformer secondary. From Sec. 13.6 it is seen that the optimum turn ratio is equal to $(R_l/R_1)^{1/2}$. At either lower or higher frequencies than that corresponding to ω'' for which the transformer is designed the effect of the associated reactances is such as to reduce the efficiency. At

low frequencies the shunt inductance kL_p reduces the potential applied to the secondary circuit. And at high frequencies the series leakage inductance L_l becomes important in reducing the potential applied to the load. The elements of Fig. 14.3 that become significant at these frequency extremes are included in the circuits of Fig. 14.4. The behavior

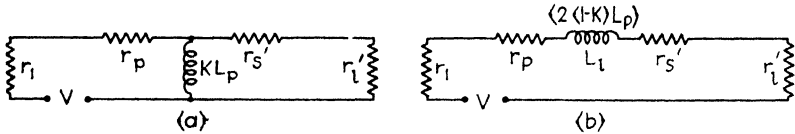



FIG. 14.4.—The equivalent circuits of the power transformer with low resistances in series with both primary and secondary. (a) Low frequencies; (b) high frequencies.

of the transformer in these frequency regions can be deduced by considering these simple circuits.

The potential transformer has a high resistance load such as the grid circuit of a vacuum tube. The actual voltage ratio is somewhat less than the ideal value n because of the primary-circuit losses. But a well-designed transformer yields a potential amplification of approximately n near the center of its intended frequency range. At low frequencies, as in the case of the power transformer, the shunt inductance

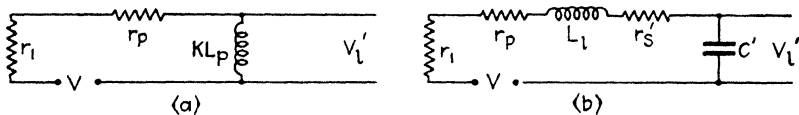



FIG. 14.5.—The equivalent circuits of a potential transformer with a high load resistance such as the grid circuit of a vacuum tube. (a) Low frequencies; (b) high frequencies.

path becomes important. The circuit is indicated in Fig. 14.5. As the frequency decreases, the reactance across which the output potential effectively appears becomes less and the output potential, V_l or nV_l', decreases. This low-frequency response can be improved by increasing the primary inductance or by placing a condenser in series with the primary. The latter forms a low-frequency resonant circuit with the primary inductance and brings up the output potential in the neighborhood of the resonant frequency. The schematic circuit and the general nature of the response for various primary Q values is indicated at the left in Fig. 14.6. It is evident that if Q is too large, an undesired hump appears. A value in the neighborhood of unity is about the optimum. L_p and R_p are fixed, the frequency region to be enhanced determines the product CL_p, and the proper primary series resistance is then determined by the chosen value of Q_1. A somewhat similar situation is inherent in the transformer itself at high frequencies. The equivalent transformer circuit is the one at the right in Fig. 14.5. The primary capacity can be neglected in general and C' is n^2 times the total dis-

tributed and load capacity across the secondary. The resonant frequency is determined by the product of this and the leakage inductance, and if the resistance of the circuit is small, a pronounced hump will appear in the characteristic at this frequency. This generally occurs in the high audio range and is an advantage in bringing up the response in that region. If a flat characteristic is desired, the Q of the circuit should be reduced to approximately unity, which is sometimes accomplished by winding the secondary with resistance wire.[1]

14.2. The Resonant Air-core Transformer.—At frequencies above the audio range, losses associated with the core become so important that

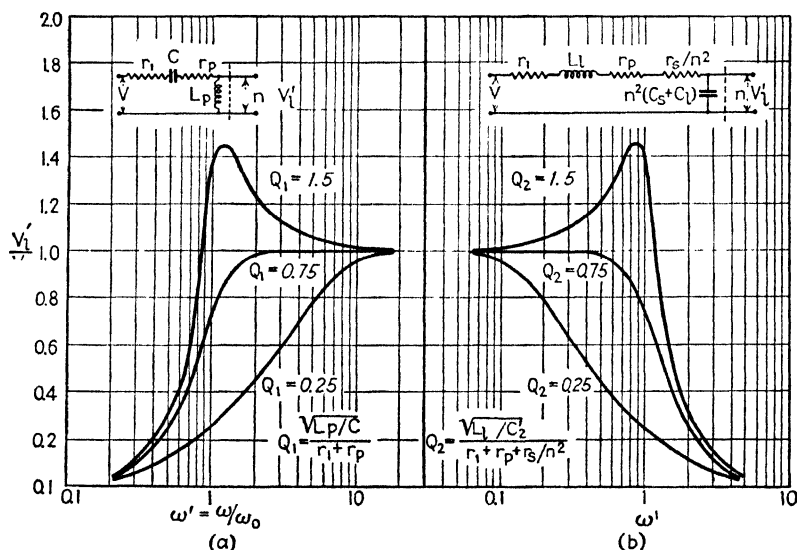
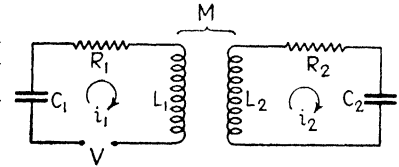



FIG. 14.6.—(a) Condenser in series with primary of a transformer to enhance low-frequency response. (b) Resonance effect of leakage reactance and secondary distributed capacity at high frequencies.

only the most carefully designed and constructed iron-core instruments are of value. In the radio-frequency range above 10^5 cycles the core is generally dispensed with entirely. The theory is fundamentally the same for the transformer without a core, but the approximations of the preceding section are no longer valid. The coupling coefficient is generally much less than unity and all associated capacities must be taken into consideration. Practically all the important radio-frequency applications involve the use of resonant or tuned primary and secondary circuits, and the present discussion will be limited to this type of circuit. These transformers perform much the same types of service as do audio-frequency ones. They isolate circuits as far as direct-current components

[1] General references: M.I.T. STAFF, "Magnetic Circuits and Transformers," John Wiley & Sons, Inc., New York, 1943; LEE, "Electronic Transformers and Circuits," John Wiley & Sons, Inc., New York, 1947.

are concerned and facilitate power transfer by matching the impedance between networks. The latter is generally accomplished by varying the mutual inductance between the primary and the secondary. In addition the selective property of the tuned circuits is made use of to discriminate against all but one narrow band of frequencies that is passed efficiently by the transformer. This is in distinction to the audio-frequency transformer which passes without discrimination a wide band on either side of the mean frequency for which it is designed. However, the absolute width of the response curve of a tuned circuit (Fig. 13.4 or 13.13) on a frequency scale is proportional to the resonant frequency. Thus, though the width of the band passed by a radio-frequency transformer may be only ± 0.5 per cent of the resonant frequency, this will include the entire audio-frequency range from 0 to 10^4 cycles if the resonant frequency of the transformer is 1 megacycle. Hence for communication purposes the discrimination is an advantage in eliminating undesired radio frequencies without unduly restricting the passage of audio-frequency modulation.

FIG. 14.7.—The air-core transformer.

The representative circuit is illustrated in Fig. 14.7. Equations (14.1) to (14.3) apply immediately to this circuit; including the transformer inductances in the circuit reactances, they become

$$\begin{aligned} \mathbf{V} &= \mathbf{z}_1\mathbf{i}_1 - j\omega M\mathbf{i}_2 \\ O &= -j\omega M\mathbf{i}_1 + \mathbf{z}_2\mathbf{i}_2 \end{aligned} \tag{14.5}$$

and

$$\mathbf{z}' = \mathbf{z}_1 + \frac{\omega^2 M^2}{\mathbf{z}_2} \tag{14.6}$$

$$\mathbf{i}_2 = \mathbf{V}\frac{j\omega M}{\mathbf{z}_1\mathbf{z}_2 + \omega^2 M^2} \tag{14.7}$$

Equation (14.6) is particularly useful in determining the impedance presented to the source of emf. The contribution made by the second term is known as the *reflected impedance* of the secondary circuit. Expressing $\mathbf{z}'$ as the sum of real and imaginary parts

$$\mathbf{z}' = \left(R_1 + \frac{\omega^2 M^2 R_2}{R_2^2 + X_2^2}\right) + j\left(X_1 - \frac{\omega^2 M^2 X_2}{R_2^2 + X_2^2}\right)$$

it is seen that the effective reactance contributed by the secondary is of the opposite sign to its true reactance. The resonant or unity-power-factor condition is that for which

$$X_1 = \frac{\omega^2 M^2}{R_2^2 + X_2^2} X_2 \tag{14.8}$$

The condition for maximum secondary current or maximum power transfer to a resistance in the secondary can be derived from Eq. (14.7). From this equation the magnitude of the secondary current is

$$i_2 = V\frac{\omega M}{\sqrt{(R_1R_2 - X_1X_2 + \omega^2M^2)^2 + (R_1X_2 + R_2X_1)^2}} \tag{14.9}$$

The determination of the maximum value of i_2 depends on which of the parameters is variable. Consider first that the secondary reactance can be varied, *i.e.*, that circuit is tunable. Setting the partial derivative of Eq. (14.9) with respect to X_2 equal to zero yields

$$X_2 = \frac{\omega^2M^2}{R_1^2 + X_1^2}X_1 \tag{14.10}$$

This means that the secondary must be tuned to a frequency different from its natural frequency in order to obtain the largest possible current. If the primary circuit alone is adjustable for a maximum i_2, the condition of Eq. (14.8) is obtained. If Eq. (14.10) holds, Eq. (14.9) becomes

$$i_2 = V\frac{\omega M(R_1^2 + X_1^2)^{1/2}}{R_2(R_1^2 + X_1^2) + \omega^2M^2R_1} \tag{14.11}$$

which is the secondary current for the optimum secondary adjustment. If the primary circuit is adjustable as well as the secondary and the secondary is continually altered so that it remains in its optimum condition, the extreme secondary current is obtained when the partial derivative of Eq. (14.11) with respect to X_1 is equal to zero. On performing this differentiation the following two conditions are obtained:

$$X_1 = 0 \qquad \text{and} \qquad X_1 = \pm\sqrt{\frac{R_1}{R_2}}\sqrt{\omega^2M^2 - R_1R_2} \tag{14.12}$$

Case 1.—$R_1R_2 > \omega^2M^2$. This is known generally as *deficient coupling* and since here the second condition of Eq. (14.12) yields an imaginary value for X_1, it has no physical significance. Only the first condition applies and the maximum value of the secondary current occurs for $X_1 = X_2 = 0$ and is given by Eq. (14.11) as

$$i_{2\max} = V\frac{\omega M}{R_1R_2 + \omega^2M^2} \tag{14.13}$$

Case 2.—As the mutual inductance is increased, for instance, by moving the coils closer together, the maximum secondary current rises until the point $R_1R_2 = \omega^2M^2$ is reached. At this point the two conditions of Eq. (14.12) coincide and the maximum current is given by

$$i_{2\max} = V\frac{1}{2\sqrt{R_1R_2}} \tag{14.14}$$

This condition is known as *critical coupling.*

Case 3.—When M has been increased so that $R_1R_2 < \omega^2M^2$, both conditions of Eq. (14.12) apply, but further analysis shows that the first yields a minimum, given of course by Eq. (14.13). The substitution of the second condition in the equation for i_2 yields maximum values which are the same as that of Eq. (14.14). The choice of positive or negative sign for the optimum reactance indicates that there are two settings for maximum current. Equation (14.10), which is assumed to be satisfied during all adjustments, implies at these maxima that $X_1/R_1 = X_2/R_2$. This case in which two current maxima exist is known as *sufficient coupling*. The efficiency of power transfer at these optimum conditions is obtained by dividing the secondary power $i_2^2R_2$ by the total power delivered to the circuit. For the condition $X_1 = X_2 = 0$ this ratio is $\omega^2M^2/(R_1R_2 + \omega^2M^2)$. It is less than $\frac{1}{2}$ for deficient coupling and equal to $\frac{1}{2}$ at critical coupling. For the greater values of ωM corresponding to sufficient coupling the maximum power delivered to the secondary at resonance remains the same, also the primary current for optimum adjustment remains $V/2R_1$, and the efficiency of power transfer continues to be $\frac{1}{2}$.

The maximum secondary current as a function of M is obtained by setting the partial derivative of Eq. (14.9) with respect to M equal to zero. This yields

$$(R_1^2 + X_1^2)(R_2^2 + X_2^2) = \omega^4M^4$$

For the optimum resonant condition in which X_1 and X_2 are both zero this equation becomes $\omega M = \sqrt{R_1R_2}$, which is the critical coupling condition. If the reactances are not equal to zero, the coupling must be greater than critical for maximum power transfer. This is merely an alternative way of looking at the preceding results. The number of parameters involved in Eq. (14.8) makes its consideration as a function of ω rather difficult. For a complete discussion of the general case reference should be made to more specialized texts.[1] The principal features are brought out by a consideration of the special case in which the circuits have the same natural frequency and the same ratio of inductance to resistance, *i.e.*, $\omega_0^2 = 1/L_1C_1 = 1/L_2C_2$ and

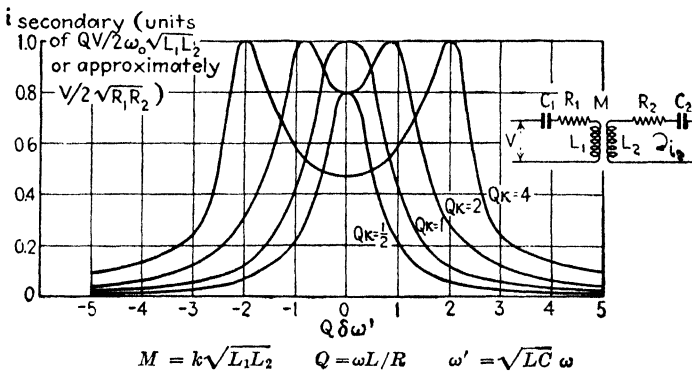
$$Q = \frac{\omega L_1}{R_1} = \frac{\omega L_2}{R_2}$$

Since Q is generally more nearly constant than R, Eq. (14.9) will be written in terms of these parameters as

[1] CHAFFEE, "Theory of Thermionic Vacuum Tubes," McGraw-Hill Book Company, Inc., New York, 1933. McILWAIN and BRAINERD, "High Frequency Alternating Currents," John Wiley & Sons, New York, 1931. TERMAN, "Radio Engineering," McGraw-Hill Book Company, Inc., New York, 1947.

$$i_2 = \frac{V}{\omega\sqrt{L_1L_2}} \frac{kQ^2}{\sqrt{[1 + (Qk)^2]^2 + 2[1 - (Qk)^2](Qx)^2 + (Qx)^4}} \tag{14.15}$$

where $k = M/\sqrt{L_1L_2}$ and $x = [1 - (\omega_0/\omega)^2]$. For values of ω in the neighborhood of ω_0, x is approximately $2\delta\omega/\omega_0$. In Fig. 14.8, i_2 is plotted in units of $QV/2\omega_0\sqrt{L_1L_2}$ as a function of $Q\delta\omega/\omega_0$ in the neighborhood of resonance for a few representative values of the parameter Qk. The critical coupling condition corresponds to $Qk = 1$. For smaller values of Qk the resonant peak appears at $\omega = \omega_0$ and for greater values two peaks occur which are of equal height on this scale and to this approximation. In practice it is difficult to achieve this symmetry and the peaks

$M = k\sqrt{L_1L_2}$ $Q = \omega L/R$ $\omega' = \sqrt{LC}\,\omega$

Maxima separated a distance

$$\delta\omega' = \frac{1}{Q}[(Qk)^2 - 1]^{1/2} \text{ for small } \delta\omega'.$$

FIG. 14.8.—Typical resonance curves for coupled systems.

are not identical in height or in displacement on either side of ω_0. However, the curves are representative of the general phenomena encountered for a variation in frequency.

14.3. Electromechanical Transducers.—In communication work the energy associated with sound waves must be converted into electrical energy, in which form it is transferred from one point to another, and then reconverted into acoustical energy. The devices for accomplishing these interconversions, such as microphones and speakers, are known generically as *transducers*. The mechanical portion of these elements is generally of a resonant type, which is an advantage for the transfer of power but a disadvantage in obtaining a uniform response independent of the frequency. In practice it is necessary to resort to various compromise designs. Also, the devices are not in general linear, *i.e.*, the mechanical displacement is not directly proportional to the circulated charge. Thus a certain amount of harmonic distortion and frequency mixing occurs during the conversion process. This disadvantage sets a practical limit to the amplitude of motion. Though the acoustical

analysis has many points in common with the preceding circuit analysis, it is beyond the scope of this discussion. For a detailed account of this theory and a description of the different types of devices that are employed the reader is referred to treatises in this field.[1] In addition to devices for the interconversion of electric and acoustic energy high-Q electromechanical systems are used for filtering. Certain substances such as quartz, which is electrostrictive, and nickel, which is magnetostrictive, have very low internal viscosities. The result is that a vibrating system composed of these materials has a very small decrement or a large effective Q value. Furthermore, the dependence of their natural frequency on the variables such as temperature and pressure can be made small, and if these are held within fairly close limits, an extreme constancy of frequency can be attained. These systems can be coupled (*i.e.*, arranged to interact) with electrical circuits and produce electrical oscillations of the desired frequency or act as filters for transmitting or suppressing this frequency.

The formal similarity between the differential equations for an oscilla, tory mechanical system and an electric circuit is very useful in discussing the combined system. Let x be a mechanical displacement and F a mechanical force, the differential equation of the mechanical system is then

$$\alpha\frac{d^2x}{dt^2} + \beta\frac{dx}{dt} + \gamma x = F \tag{14.16}$$

The Greek letters stand for the parameters of the system and depend on the particular type of motion being represented. In the case of linear motion α is the effective mass, β is the retarding force per unit velocity, and γ is the restoring force per unit linear displacement. Representative rotational motion coefficients are those of Eq. (10.9). If q is the charge on a condenser in an electrical circuit and V an applied electromotive force, the circuit equation is

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = V \tag{14.16'}$$

As they stand, these two equations are quite independent, but the systems that are under consideration are those for which there is some electromechanical interaction. In this case the mechanical force F contains a term which depends on one of the circuit vectors and the emf. V contains a term which in general depends on the mechanical displacement or

[1] RAYLEIGH, "Theory of Sound," 2d ed., The Macmillan Company, New York, 1926; DAVIS, "Modern Acoustics," The Macmillan Company, 1934; MORSE, "Vibration and Sound," McGraw-Hill Book Company, Inc., New York, 1948; OLSEN, "Acoustics," D. Van Nostrand Company, Inc., New York, 1947; MASON, "Electromechanical Transducers and Wave Filters," D. Van Nostrand Company, Inc., New York, 1942.

velocity. Assume that s_m represents a mechanical restoring force per unit charge q and s_e an electrical restoring force, or reverse emf., per unit mechanical displacement. Further assuming a constancy of the coefficients and that in the steady state the mechanical and electrical vectors are simple periodic functions of the time, the equations become

$$(-\alpha\omega^2 + j\omega\beta + \gamma)\mathbf{x} + s_m\mathbf{q} = \mathbf{F}$$
$$s_e\mathbf{x} + \left(-L\omega^2 + j\omega R + \frac{1}{C}\right)\mathbf{q} = \mathbf{V}$$

in the complex notation. It is generally more convenient to write these in terms of the current, $\mathbf{i} = d\mathbf{q}/dt$, and the mechanical velocity, $\mathbf{v} = d\mathbf{x}/dt$. Differentiating the equations with respect to t and dividing through by $j\omega$, they may be written

$$\begin{aligned} \mathbf{z}_{11}\mathbf{v} + \mathbf{z}_{12}\mathbf{i} &= \mathbf{F} \\ \mathbf{z}_{21}\mathbf{v} + \mathbf{z}_{22}\mathbf{i} &= \mathbf{V} \end{aligned} \tag{14.17}$$

where $\mathbf{z}_{11}$ is the mechanical impedance, $\mathbf{z}_{22}$ the electrical impedance, and $\mathbf{z}_{12}$ and $\mathbf{z}_{21}$ the interaction factors $s_m/j\omega$ and $s_e/j\omega$. These equations are very similar to those for two coupled electrical circuits. They are simple linear equations which may be solved for the current or velocity in terms of the applied forces. The solutions, of course, are similar to those for the air-core transformer. This discussion has been limited to the case of two simple periodic circuits, one mechanical and one electrical. The extension to more complex cases is similar to the extension of the general circuit theory of Sec. 13.6. This, however, introduces no new principles and will not be considered further. The chief points of interest are those associated with the interaction factors and the mechanical impedance and these are best illustrated by a few specific examples.

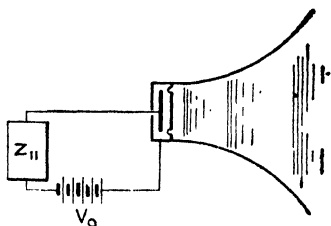



FIG. 14.9.—Electrostatic-acoustic system.

Both microphones and speakers can be constructed with piezoelectric crystals or on the simple condenser principle. In the case of the condenser the arrangement is schematically that of Fig. 14.9. The formed baffle or horn acts as an acoustic transformer coupling the diaphragm to the atmosphere. The diaphragm is one plate of the condenser and in general it does not vibrate as a piston but as a clamped plate and the effective mass, retarding, and restoring forces are functions of the frequency and rather difficult to calculate. The power consumed by the mechanical system is $v^2\beta$. A considerable

fraction of this is radiated as sound in the case of the speaker. To determine the electroacoustical efficiency the mechanical parameters must be measured as a function of the frequency and when the mechanical power has been obtained, this must be multiplied by the efficiency of the power transfer from the mechanical to the acoustical form to obtain the acoustic energy radiated. The energy stored in the condenser, which is the electroacoustical element, is $q'^2/2C$, where q' is the charge and C is the capacity. This can be written as $\frac{q'^2(d + x)}{2C'}$, where C' is written for $C(d + x)$ which for values of the displacement x, small in comparison with the total separation d, can be taken as a constant Cd. The mechanical force is the negative partial derivative of this with respect to x, which is $-\frac{q'^2}{2C'}$. The charge q' is made up of a large constant charge q_0 due to the direct-current potential in the circuit plus a small periodic charge q. Expanding q'^2 and neglecting the small term in q^2, the periodic mechanical force is seen to be $-\frac{q_0 q}{C'}$, or the mechanical force per unit charge s_m is $-\frac{q_0}{C'} = -\frac{V_0}{d}$. The negative partial derivative of the energy with respect to the charge is the negative of the potential or restoring potential. On forming this quantity and dividing the periodic term by the displacement x the electrical force per unit displacement is seen to be $-\frac{V_0}{d}$ also. Thus $\mathbf{z}_{12} = \mathbf{z}_{21} = jV_0/\omega d$, and the interaction factor is proportional to the direct-current potential V_0 and inversely proportional to the separation d. The latter cannot be decreased indefinitely since it must be kept large in comparison with the maximum amplitude of motion of the diaphragm. Comparing this with the transformer coupling coefficient $-j\omega M$, it is seen to have the opposite sign and that ω enters in the denominator instead of the numerator. The former merely indicates a change of phase π, and the latter means that the coupling decreases with the frequency instead of increasing as in the case of the transformer.

Electromagnetic principles can also be used in the construction of microphones and speakers. A larger amount of power can be handled by this type of device and it is used almost exclusively in power speaker elements. The headphone, illustrated in Fig. 14.10, is representative of a low-power element. The coil of the electromagnetic circuit is generally composed of a large number of turns of very fine wire since the resistance is not a great disadvantage and a large flux is produced for a small current. The magnetic circuit is composed of a permanently

polarized armature which is separated by a small air gap from the iron diaphragm. The permanent polarization of the armature should be large since the permanent flux plays the same role here as does the constant potential in the condenser instrument. Similarly, the air gap should be small, subject of course to the restriction imposed by the maximum amplitude of vibration. The hysteresis loss and nonlinearity of the magnetic circuit are disadvantages which are invariably associated with this type of device. Neglecting the fringing of flux between the armature pole and diaphragm, the force of attraction between the two is $\frac{1}{2}\phi'^2/\mu_0 A$, where A is the area of the pole face. Expanding this in terms of the permanent flux ϕ_0 and the periodic flux ϕ, neglecting ϕ^2, yields a constant minus the periodic term $\phi_0\phi/\mu_0 A$, or since $\phi = ni/\mathcal{R}$, where n is the number of turns in the coil, the periodic force per unit current, or $\mathbf{z}_{12}$, is $-\frac{\phi_0 n}{\mu_0 \mathcal{R} A}$. The emf. induced by a change in x is $-n\frac{\partial\phi'}{\partial t}$ and since ϕ' changes through the reluctance, $\mathcal{R}$, this is $\phi_0 n/\mu_0\mathcal{R}A$ times dx/dt, or v. Thus the emf. per unit velocity, which is $\mathbf{z}_{21}$, is $-\mathbf{z}_{12}$. From this it is seen that for a large coupling the constant flux and n should be large and the reluctance of the magnetic circuit should be small. Since $\mathcal{R}$ is approximately equal in magnitude to $d/\mu_0 A$, where d is the gap between the armature and diaphragm, the interaction factors can be written $\mathbf{z}_{21} = -\mathbf{z}_{12} = \varphi_0 n/d$. The difference in sign is a consequence of Lenz's law and characteristic of all electromagnetic instruments. It represents a phase shift of π between the mechanical force and the current produced by it, but since j does not appear in the $\mathbf{z}$'s, the input or output impedance of the device as calculated from Eqs. (14.17) retains its same form.

FIG. 14.10.—Headphone type of acoustical element.

FIG. 14.11.—Schematic diagram of a dynamic speaker.

For larger power devices the headphone type of instrument is unsuitable and the so-called dynamic speaker is employed. The principle of this instrument is illustrated in Fig. 14.11. A constant magnetomotive force is produced by the shaded coil and results in a large radial induction through the narrow annular ring. In this ring is a short, light solenoid through which the alternating current flows. It is rigidly connected to the apex of a semiflexible fiber cone which acts as the source of the sound waves. This does not vibrate as a simple piston, but the type of motion

is a function of the applied frequency. Thus the effective mass, acoustical resistance, and restoring force vary throughout the audible range and it is difficult to design a speaker that is entirely satisfactory over the complete spectrum. If B_0 is the induction in the gap and l the total length of wire, the axial force per unit current on the cone is B_0l. This is equal in magnitude but opposite in sign to the emf. per unit velocity, hence $\mathbf{z}_{12} = -\mathbf{z}_{21} = B_0l$. From this it is evident that the amplitude of motion does not limit the linearity of response.

If the interaction factors are known, the effective mechanical parameters α, β, and γ corresponding to any one mode of oscillation can be determined by measuring the electrical impedance of the device as a function of the frequency. If the moving element is clamped, the input impedance, $\mathbf{V}/\mathbf{i}$, is equal to $\mathbf{z}_{22}$. The input impedance when unclamped is $\mathbf{z}_{22} + \frac{A^2}{\mathbf{z}_{11}}$, where A^2 is written for the product of the interaction factors.

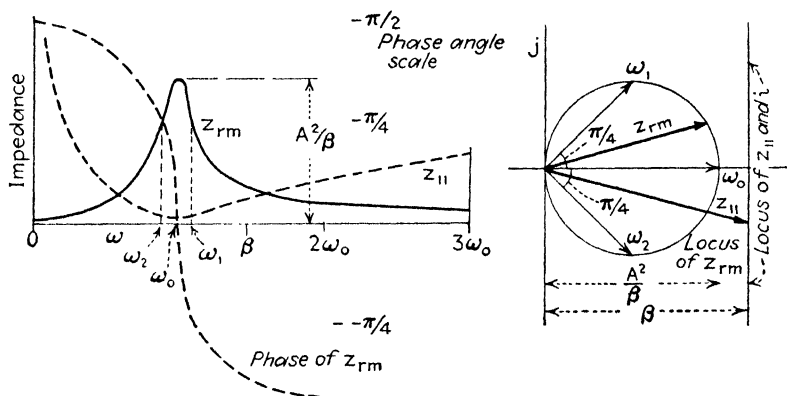



FIG. 14.12.—Analysis of a resonant electroacoustical system.

The difference between the impedance unclamped and clamped is the *reflected mechanical impedance* $\mathbf{z}_{rm} = A^2/\mathbf{z}_{11}$. The locus of the mechanical impedance vector $(\beta + jX_{11})$ as a function of the frequency is a straight line parallel to the imaginary axis intersecting the real axis at β. If A is independent of the frequency, as for electromagnetic devices, $\mathbf{z}_{rm}$ is inversely proportional to $\mathbf{z}_{11}$ and hence its locus in the complex plane is a circle (see circle diagrams, Sec. 13.4). These relations are shown in Fig. 14.12. The constant β can be determined from the fact that at resonance the magnitude of the effective resistance is A^2/β. If ω_0 is the angular frequency of resonance, the ratio of α to γ is given by

$$\omega_0 = \sqrt{\gamma/\alpha}$$

The third relation which is necessary to determine the three quantities can be obtained most conveniently from the angular frequencies on

either side of resonance for which the magnitude of the reflected impedance has fallen to $1/\sqrt{2}$ of its resonant value. From the diagram these are seen to correspond to phase angles of $\pi/4$ and $-\pi/4$, *i.e.*,

$$\frac{\alpha\omega_1 - \gamma/\omega_1}{\beta} = 1 \qquad \frac{\alpha\omega_2 - \gamma/\omega_2}{\beta} = -1$$

Eliminating γ, $\omega_1 - \omega_2 = \beta/\alpha$. These relations determine all the mechanical parameters in terms of the resistance at resonance R, and A, ω_0, ω_1, and ω_2.

$$\alpha = \frac{A^2}{R(\omega_1 - \omega_2)} \qquad \beta = \frac{A^2}{R} \qquad \gamma = \frac{A^2\omega_0^2}{R(\omega_1 - \omega_2)}$$

If A is not known, it can be determined experimentally by making a simultaneous electrical and mechanical measurement. For instance, if V' is equal to the difference between the potential across the element at resonance when clamped and when unclamped, from Eq. (14.17) $V' = Av$, since A is the magnitude of $\mathbf{z}_{21}$. If the amplitude of motion at resonance, say x_0, is also measured, $v = \omega_0 x_0$, or $A = V'/\omega_0 x_0$. The circle diagram is very convenient for the preceding type of analysis, but it does not bring out the variation in magnitude and phase of the impedance with frequency as clearly as the diagram at the left in Fig. 14.12. This is representative of a circuit with a fairly high decrement or low Q_0 value. As has been seen in previous sections, Q_0 may be considered as a measure of the sharpness of the resonant peak. The mechanical Q_0 is $\omega_0\alpha/\beta$ or $\omega_0/(\omega_1 - \omega_2)$.

FIG. 14.13.—Double-button carbon microphone.

Another type of microphone which is widely used is the carbon button type illustrated schematically in Fig. 14.13. A small cup or button containing carbon granules is in contact with a metal diaphragm in such a way that a motion of the diaphragm changes the pressure on the carbon grains. These form part of an electric circuit and the change in pressure results in a change of resistance. If R is the mean resistance of the circuit and r the change in resistance due to the motion of the diaphragm, the current in the circuit is given by

$$i = \frac{V}{R + r} = \frac{V}{R}\left[1 - \frac{r}{R} + \left(\frac{r}{R}\right)^2 \cdots\right]$$

Thus the device is not linear when operated at constant potential and harmonics are introduced. These can be minimized by keeping the mean resistance R very large in comparison with r. Even harmonics are eliminated by the double-button type, which is the one actually shown in the figure. The varying r is of the opposite sign for the two circuits and the output current from the transformer is proportional to

the difference between i_1 and i_2, hence from the preceding equation

$$i = \frac{2V}{R}\left[\frac{r}{R} + \left(\frac{r}{R}\right)^3 \cdots\right]$$

Thus the even harmonics and the constant term are eliminated. The removal of the direct current is particularly advantageous, for it would polarize the transformer core and reduce its effective permeability. Microphone sensitivity is given in terms of decibels below an arbitrary standard. This is generally chosen as the production of an open circuit emf. of 1 volt for a pressure of 1 dyne per square centimeter (bar). (The threshold of hearing varies from a sound-wave pressure of 10^{-4} bar in the neighborhood of a few thousand cycles to about 1 bar at very high or very low frequencies.) Thus a −30-db. microphone would require to be followed by a 30-db. amplifier for the least audible sounds to produce an emf. of 1 volt.

An efficient acoustical element is characterized by a large damping coefficient. This implies a large power transfer from the electrical circuit and a minimum of frequency selectivity. The other group of electromechanical elements for generating particular frequencies or for

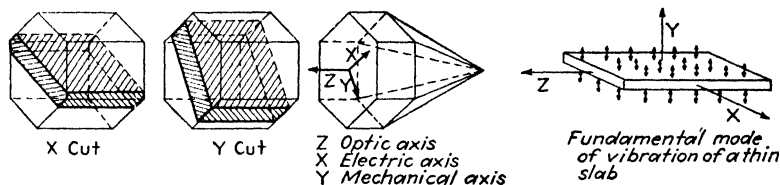



FIG. 14.14.—Quartz-crystal resonator.

discriminating between frequencies that are very close together is characterized by a very small damping factor. The piezoelectric quartz crystal, which is representative of this type, will be discussed briefly. The main portion of the ideal quartz crystal is hexagonal in cross section, as indicated in Fig. 14.14. The Z axis perpendicular to a hexagonal section is known as the optic axis. If plane-polarized light (Sec. 16.2) is sent through the crystal along this axis, it is found that the plane of polarization is rotated about the direction of propagation. The X, or electric, axis is perpendicular to Z and in the direction of an apex of the hexagon. The third axis, designated by Y, is sometimes known as the mechanical axis and is perpendicular to a side of the hexagon. The crystal shows piezoelectric properties in any direction except along the optic axis. The crystals used in electric circuits are generally in the form of thin slabs cut parallel to the XZ or YZ planes. These are known as X or Y cuts, respectively. Other cuts are also made for obtaining special advantages such as low-temperature coefficients or particularly stable modes of oscillation.[1] In the fundamental mode of oscillation

[1] For a complete discussion of the preparation of quartz crystals and their use in electric circuits see Vigoureux, Quartz Resonators and Oscillators, *Nat. Phys. Lab.* (1931); Mason, *Bell. System Tech. J.*, **13**, 405 (1934); Lack, *Bell. System Tech. J.*, **13**, 453 (1934); CADY, "Piezoelectricity," McGraw-Hill Book Company, Inc., New York, 1946.

such a slab expands and contracts along an axis normal to the principal faces. The central plane remains at rest and the two faces have the maximum amplitude of motion. The thickness of the crystal is then one-half the length of a mechanical wave in the medium. The velocity of propagation is equal to the square root of the elasticity over the density and is also equal to the product of the frequency and wave length. Therefore, if t is the thickness of the specimen, E its elasticity, and ρ its density, the frequency ν of the fundamental mode is given by

$$\nu = \frac{1}{2t}\sqrt{\frac{E}{\rho}}$$

It is found that the elasticity is a function of the direction in the crystal. For an X cut the product νt in megacycle millimeters is 2.86, and for a Y cut it is 1.96. Thus the thickness and the orientation of the slab with respect to the crystal axes determine the natural frequency. There is a small temperature coefficient of natural frequency. For the X cut, which is the one in most general use, it is -22 cycles per megacycle per degree

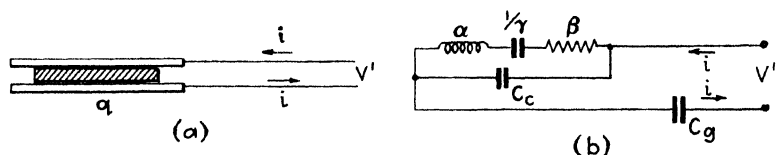



FIG. 14.15.—(a) Quartz crystal between metal plates as part of an electric circuit. (b) Equivalent electrical circuit.

centigrade. Thus thermostating to 0.1°C. will keep the natural frequency constant to 2 parts in 10^6.

When included in an electric circuit, the crystal is mounted between two metal plates, as shown schematically in Fig. 14.15. The lower portion of the figure represents the equivalent electrical circuit. If there is an air gap above the crystal, its capacity is represented by C_g, and C_c is the pure electrostatic capacity of the holder and crystal. The Greek letters represent the equivalent crystal parameters. Since the piezoelectric constant is 6.4×10^{-8} esu. per square centimeter per dyne per square centimeter and the elastic constant is 7.85×10^{11} dynes per square centimeter the charge q' in coulombs per square meter for a displacement x is given by $q' = 0.167x/t$. The charge on one side of the crystal is equal to q' times A, the area of the face. In terms of current and velocity $v = 6ti/A$. For most purposes, however, no particular interest attaches to the mechanical motion of the crystal. At resonance the crystal moves in phase with the applied electric force and the velocity and current are large. The frequency of resonance is determined by the cut and thickness and the magnitude of the current at resonance is inversely proportional to β; the latter is proportional to

the crystal viscosity, which is small, hence the resonant current is large. The ratio $\omega_0\alpha/\beta$ is the Q_0 value of the crystal as a resonant electrical element. It is larger by a factor of from 10 to 100 than for the best electrical circuits. The practical frequency limits of the oscillating crystal element are from about 50 kilocycles to 15 megacycles. The Q_0 value in the neighborhood of 0.1 megacycle is of the order of 10^4, which means that a change of only 5 cycles in the applied frequency is enough to change the phase angle of the impedance presented by the crystal by $\pi/4$. This results in a decrease in current magnitude by a factor of $1/\sqrt{2}$ from the resonant value. The very rapid change of impedance with frequency makes crystal elements valuable for suppressing or transmitting narrow frequency bands. This characteristic is also employed for controlling the frequency of an oscillator, as will be seen in Sec. 15.6.

14.4. Frequency or Wave Filters.—For many purposes it is desirable that the electrical circuit should freely transmit certain regions of the frequency spectrum and completely suppress others. In certain direct-current systems it is an advantage to suppress completely all alternating-current components. Thus series inductances and shunt capacities are used following a rectifier to eliminate the power-frequency component and its harmonics. In audio-frequency communication work it is necessary to suppress power-frequency hums and high carrier and radio frequencies that may be present. And in many radio-frequency applications it is desirable to suppress all but a very narrow frequency band. The frequency discrimination of the two types of reactance and their simple series and parallel combinations has been discussed in the preceding chapter, and the use of piezoelectric elements has just been mentioned. The ideal element would be one which presented no impedance for the desired frequencies and an infinite one for all others. This ideally sharp cutoff is not attainable. But just as two resonant systems following one another in cascade have a higher discrimination than a single one, a series of meshes can be assembled which approximate the ideal filter. The principles involved in the design of this type of network are of general interest and one or two specific examples will also be discussed to illustrate the type of frequency discrimination that can be attained.

The type of filter network most commonly encountered is a cascade assemblage of units, consisting of series and parallel elements. It is known as a passive quadripole, for it has two input and two output terminals and contains no internal sources of emf. Such a network is illustrated schematically in Fig. 14.16. There are certain general circuit theorems concerning it that are of interest. Assume that there is a potential difference $\mathbf{V}_1$ across the input terminals and a potential difference $\mathbf{V}_2$ across the output terminals. The general circuit equations [Eq.

(13.20)] may be solved for the input and output currents, yielding

$$\begin{aligned} \mathbf{i}_1 &= \frac{\mathbf{A}_{11}}{\mathbf{D}}\mathbf{V}_1 + \frac{\mathbf{A}_{12}}{\mathbf{D}}\mathbf{V}_2 \\ \mathbf{i}_2 &= \frac{\mathbf{A}_{12}}{\mathbf{D}}\mathbf{V}_1 + \frac{\mathbf{A}_{22}}{\mathbf{D}}\mathbf{V}_2 \end{aligned} \tag{14.18}$$

where **D** is the impedance determinant of the current coefficients and the **A**'s are the cofactors of the impedances with similar subscripts in this determinant. These are the simple equations for a two-mesh network and this result shows that any linear passive quadripole, no

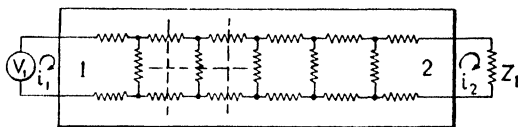



FIG. 14.16.—Passive quadripole formed by impedances in cascade.

matter how complex, can be represented formally by three impedances. The coefficients of the emfs. can be determined by the obvious measurements at the terminals and are generally written

$$\frac{\mathbf{D}}{\mathbf{A}_{11}} = \mathbf{z}_{1s}, \text{ short-circuit input impedance}$$

$$\frac{\mathbf{D}}{\mathbf{A}_{21}} = \frac{\mathbf{D}}{\mathbf{A}_{12}} = \mathbf{z}_{ts}, \text{ short-circuit transfer impedance}$$

$$\frac{\mathbf{D}}{\mathbf{A}_{22}} = \mathbf{z}_{2s}, \text{ short-circuit output impedance}$$

The input impedance with a load $\mathbf{z}_l$ across the output terminals can be determined from the compensation theorem by writing $-\mathbf{z}_l\mathbf{i}_2$ for $\mathbf{V}_2$. This yields the input impedance $\mathbf{z}'$ as

$$\mathbf{z}' = \frac{\mathbf{z}_{1s}(\mathbf{z}_{2s} + \mathbf{z}_l)}{\mathbf{z}_{2s} + \left(1 - \frac{\mathbf{z}_{1s}\mathbf{z}_{2s}}{\mathbf{z}_{ts}^2}\right)\mathbf{z}_l} \tag{14.19}$$

The impedance as viewed from the terminals marked 2 is a similar expression with the subscripts 1 and 2 interchanged. The *open-circuit impedances*, $\mathbf{z}_{10}$ and $\mathbf{z}_{20}$, are the values of these input and output impedances when the far terminals are opened. Letting $\mathbf{z}_l$ become infinite, Eq. (14.19) becomes

$$\mathbf{z}_{10} = \frac{\mathbf{z}_{1s}}{1 - \frac{\mathbf{z}_{1s}\mathbf{z}_{2s}}{\mathbf{z}_{ts}^2}} \tag{14.20}$$

From this equation and the analogous one for $\mathbf{z}_{20}$ it is evident that

$\mathbf{z}_{1s}/\mathbf{z}_{10} = \mathbf{z}_{2s}/\mathbf{z}_{20}$. These various expressions are all useful in dealing with this type of network. The input impedance for which $\mathbf{z}'$ is equal to the impedance of the terminating load $\mathbf{z}_l$ is known as the *iterative impedance*. The impedances which simultaneously terminate the two ends of the network in such a way that the impedance viewed in either direction from terminals 1 is the same and the impedance viewed in either direction from terminals 2 is the same, *i.e.*, the impedances which simultaneously match the network terminals are known as the *image impedances*. On applying this condition to Eq. (14.18) and its analogue for terminals 2, the image impedances are found to be $\mathbf{z}_{1i} = \sqrt{\mathbf{z}_{1s}\mathbf{z}_{10}}$ and $\mathbf{z}_{2i} = \sqrt{\mathbf{z}_{2s}\mathbf{z}_{20}}$ for terminals 1 and 2, respectively.

The design of filters is greatly facilitated if they can be assumed to work between their image impedances. Any number of sections may be placed in cascade, and provided the impedances are matched at intermediate junctions the over-all image impedances are given by the above expressions for the terminating sections. If the filter is terminated in its image impedances, the ratio of the volt-ampere product entering the network to that leaving it can be put in a very simple form. Substituting the load drop $-\mathbf{z}_{2i}\mathbf{i}_2$ with negative sign for $\mathbf{V}_2$ in Eqs. (14.18) and solving for $\mathbf{z}_{1i}\mathbf{i}_1^2/\mathbf{z}_{2i}\mathbf{i}_2^2$, which will be written $e^{2\theta}$

$$\tanh \theta = \frac{e^{2\theta} - 1}{e^{2\theta} + 1} = \sqrt{\frac{\mathbf{z}_{1s}}{\mathbf{z}_{10}}} \tag{14.21}$$

θ is the *image-transfer constant* and its value determines the ratio of the input volt-ampere product to the output volt-ampere product. Because of the exponential definition of θ the image-transfer constant for a series of matched sections in cascade is equal to the sum of the constants for the separate elements. θ is in general complex and its real part is known as the *image-attenuation constant* and its imaginary part as the *image-phase constant*. The same additive relation obviously holds for these quantities as well.

Of course, the impedances are usually functions of the frequency and it is not in general possible so to terminate the quadripole that impedances are matched at all frequencies. However, this condition will be assumed to a first approximation and it will further be assumed that the filter is constructed of such high Q elements that the resistance can be neglected to this approximation.[1] The most common configuration is the ladder-type net illustrated in Fig. 14.16. In the simplest case this is made up of identical symmetrical recurrent elements of the T- or π-section types of Fig. 14.17. The input and output image and

[1] For discussions not subject to these restrictions the reader is referred to Shea, "Transmission Networks and Wave Filters," D. Van Nostrand Company, Inc., New York, 1929; Guillemin, "Communication Networks," John Wiley & Sons, Inc., New York, 1931.

iterative impedances are all equal and they can be expressed in terms of the shunt and series impedances of the elementary sections. Thus

$$\mathbf{z}_{1i} = \mathbf{z}_{2i} = \sqrt{\mathbf{z}_1\left(\mathbf{z}_2 + \frac{\mathbf{z}_1}{4}\right)} \qquad \text{(T section)}$$

$$\mathbf{z}_{1i} = \mathbf{z}_{2i} = \mathbf{z}_2 \sqrt{\frac{\mathbf{z}_1}{\left(\mathbf{z}_2 + \frac{\mathbf{z}_1}{4}\right)}} \qquad (\pi \text{ section})$$

and

$$\tanh \theta = \frac{\sqrt{\mathbf{z}_1(\mathbf{z}_1 + 4\mathbf{z}_2)}}{\mathbf{z}_1 + 2\mathbf{z}_2} \qquad \text{(either type of section)}$$

The T section represents a mid-series severing and its iterative impedance is known as the *mid-series iterative impedance.* That of the π section

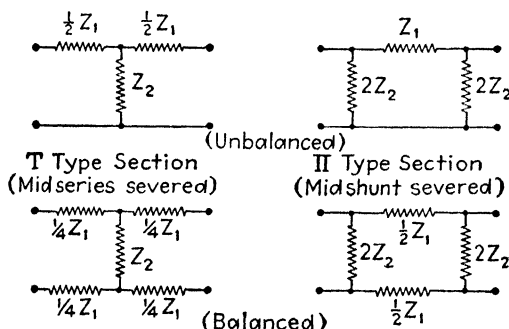



FIG. 14.17.—Typical elementary sections of a ladder-type filter.

is known as the *mid-shunt iterative impedance.* By means of the hyperbolic identities the transfer constant can be expressed more briefly as

$$\theta = \cosh^{-1}\left(1 + \frac{\mathbf{z}_1}{2\mathbf{z}_2}\right) = 2\cosh^{-1}(1 + x)^{1/2} \qquad (14.22)$$

where x is written for $\mathbf{z}_1/4\mathbf{z}_2$. If the series and shunt components are pure reactances, x is real. If x is greater than 0, there is a real solution for $\theta/2$, hence writing $\theta = \alpha + j\beta$, where α is the attenuation constant and β the phase constant

$$\alpha = 2\cosh^{-1}(1 + x)^{1/2} = 2\sinh^{-1} x^{1/2}$$
$$\beta = 0 \qquad (14.23)$$

If, however, x lies between 0 and -1, $\cosh \frac{\theta}{2}$ is a real fraction, and since

$\cosh\left(\frac{\alpha}{2} + j\frac{\beta}{2}\right) = \cosh\frac{\alpha}{2}\cos\frac{\beta}{2} + j\sinh\frac{\alpha}{2}\sin\frac{\beta}{2} = (1 + x)^{1/2}$ for which

there is a solution only if $\sinh \frac{\alpha}{2} = 0$, we have

$$\begin{aligned} \alpha &= 0 \\ \beta &= 2 \cos^{-1} (1 + x)^{1/2} = 2 \sin^{-1} (-x)^{1/2} \end{aligned} \qquad (14.24)$$

For values of x less than -1, $\cosh \frac{\theta}{2}$ is a pure imaginary, *i. e.*, $\cos \frac{\beta}{2}$ must vanish, or

$$\begin{aligned} \alpha &= 2 \sinh^{-1} (-1 - x)^{1/2} \\ \beta &= \pm\pi \end{aligned} \qquad (14.25)$$

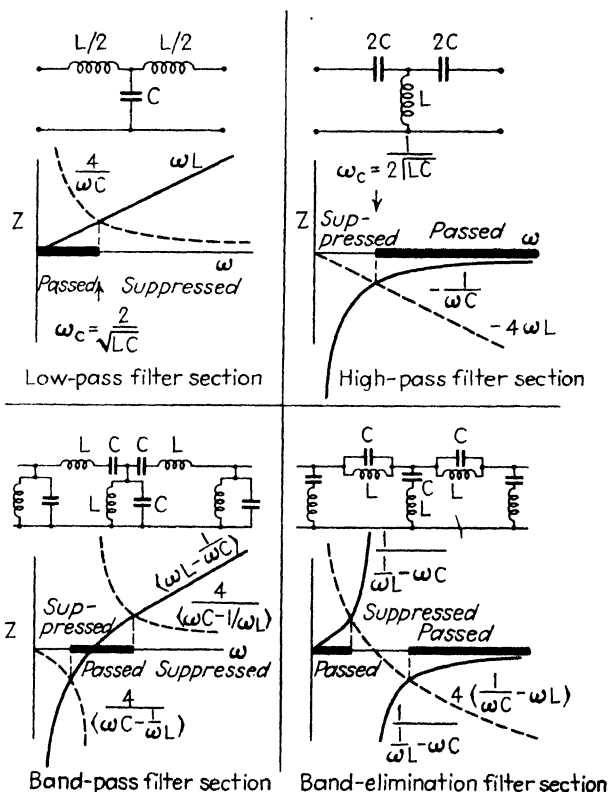



FIG. 14.18.—Frequency regions passed and suppressed by simple filter sections.

Thus the transfer constant has three distinct characters, depending on the frequency range. From the definition of this constant e^θ is the ratio of the input to output current when the filter is terminated by its image impedance or $\mathbf{i}_2/\mathbf{i}_1 = e^{-\theta} = e^{-\alpha - j\beta}$. From Eqs. (14.23) and (14.25) α is real and positive for frequencies such that $x > 0$ or $x < -1$, *i.e.*, there is a decrease in magnitude of the current or an attenuation in these ranges. In the first range there is no change in phase between

the output in input currents and in the second there is a change of π. If, however, x lies between 0 and -1, there is no attenuation and the currents are of the same magnitude, but there is a change of phase which varies throughout the range.

Figure 14.18 illustrates the graphical method of determining the frequency range passed with no attenuation and that attenuated or suppressed. Of course, the latter is not completely suppressed but the

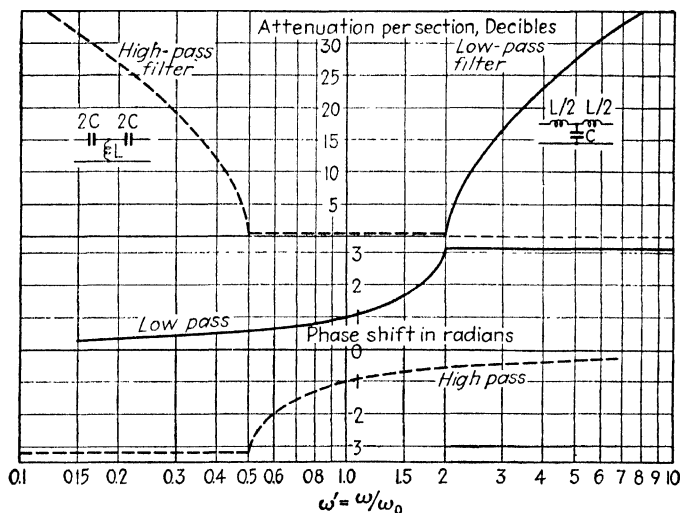



FIG. 14.19.—Attenuation and phase shift due to high- and low-pass filters.

output-input current ratio is $e^{-n\alpha}$ after n sections and this can be made very small even for frequencies close to the cutoff point if n is sufficiently large. The condition for no attenuation is

$$0 > \frac{\mathbf{z}_1}{4\mathbf{z}_2} > -1 \qquad (14.26)$$

If the impedances are of the same sign there is attenuation for all frequencies as this condition can not be satisfied. Assume they are of opposite sign and plot $\mathbf{z}_1$ and $-4\mathbf{z}_2$ as a function of the frequency. If the curves lie on the same side of the zero axis, the sign condition is fulfilled, and if $\mathbf{z}_1$ lies nearer this axis than $4\mathbf{z}_2$, the ratio is a fraction. Hence, for the frequency ranges in which this condition is satisfied the filter transmits, and for all others it attenuates. The low-pass filter has series inductances and shunt capacities. If the inductances and capacities are as shown in the diagram, angular frequencies below $2/\sqrt{LC}$ are passed and higher ones are attenuated, as shown in Fig. 14.19. The high-pass filter has series capacities and shunt inductances passing frequencies above $1/(2\sqrt{LC})$ and attenuating those below as shown

in Fig. 14.19. The simplest type of band-pass-filter section is also shown. From Eq. (14.26) the bounding frequencies are seen to be $(\sqrt{2} \pm 1)\omega_0$. ω_0 is written for $1/\sqrt{LC}$ and is seen to be the geometric mean of the bounding frequencies. The attenuation outside this band is shown in Fig. 14.20. Finally the correspondingly simple band-attenuation filter is shown. For this network Eq. (14.26) gives the bounding frequencies as $(\sqrt{17} \pm 1)\omega_0/4$. The attenuation inside the band is shown in Fig. 14.20.

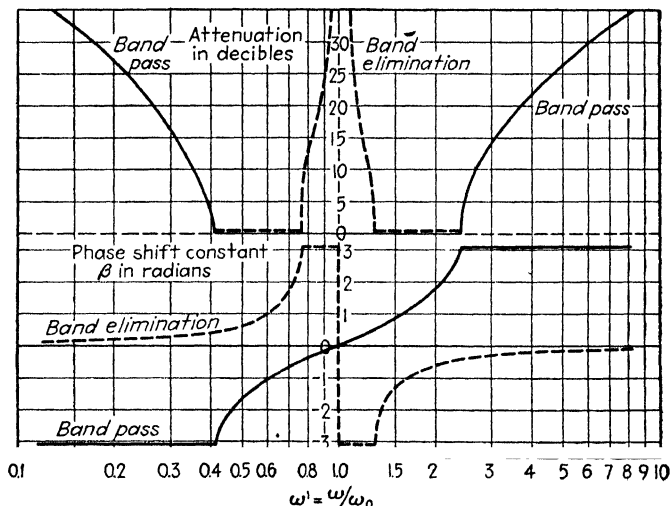



FIG. 14.20.—Attenuation and phase shift in simple band-pass and band-elimination filters.

14.5. Distributed Parameter Circuits: Lines.—The two-conductor transmission line for power or communication circuits is a four-terminal device to which the passive quadripole analysis is applicable. However, the physical length of these lines is so great that it is frequently not legitimate to consider that the current or potential distribution along them is uniform at any instant. This means that the inductance and capacity associated with them cannot be considered as localized at certain specific points but is distributed in general uniformly between the terminals. Thus as in the case of high-frequency inductances these are distributed parameter systems. They are not characterized by one natural frequency but by a fundamental frequency and all of its harmonics. Though the quadripole analysis can be extended directly to this type of system, it is more instructive to consider the propagation of alternating currents along lines beginning with the fundamental principles.

The inductance associated with the line is due to the magnetic flux that threads the long narrow loop constituting the circuit. The

capacity is the electrostatic capacity between the conductors. There is also effective resistance which includes straight ohmic resistance, skin effect, and losses due to hysteresis and eddy currents in neighboring conductors as well as dielectric losses in neighboring insulators. Finally there may also be leakage conductance between the conductors owing to imperfect insulation or corona. These quantities are generally expressed in terms of a unit loop length of the wire meaning a unit length of the conductor pair. This is illustrated schematically in Fig. 14.21. R' and L' are the resistance and inductance per unit loop length and C' and G' are the shunt capacity and conductance per unit loop length. Let $\mathbf{V}$ be the potential difference between the two conductors and $\mathbf{i}$

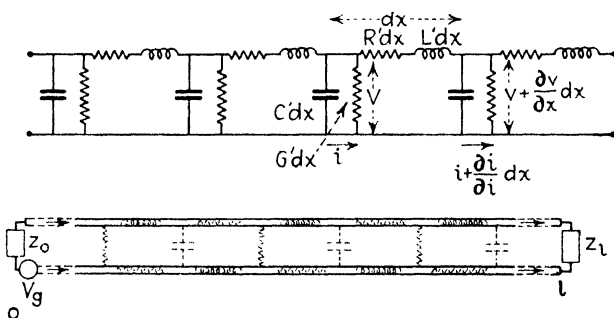



FIG. 14.21.—Schematic diagram illustrating line parameters.

be the current flowing in each at a point x along the line. The potential difference between the lines at the neighboring point $x + dx$ is

$$\mathbf{V} + \left(\frac{\partial \mathbf{V}}{\partial x}\right)dx$$

by Taylor's theorem and it is also equal to $\mathbf{V}$ minus the resistive and reactive drop which may be written $\left(R' + L'\frac{\partial}{\partial t}\right)\mathbf{i}\,dx$.

$$\mathbf{V} - \left[\mathbf{V} + \left(\frac{\partial \mathbf{V}}{\partial x}\right)dx\right] = \left(R' + L'\frac{\partial}{\partial t}\right)\mathbf{i}\,dx$$

Also the current flowing in the line at the point $x + dx$ is $\mathbf{i} + \left(\frac{\partial \mathbf{i}}{\partial x}\right)dx$ by Taylor's theorem and it is also equal to $\mathbf{i}$ minus the conductive and capacitative current between the lines in the interval dx which can be written $\left(G' + C'\frac{\partial}{\partial t}\right)\mathbf{V}\,dx$, or

$$\mathbf{i} - \left[\mathbf{i} + \left(\frac{\partial \mathbf{i}}{\partial x}\right)dx\right] = \left(G' + C'\frac{\partial}{\partial t}\right)\mathbf{V}\,dx$$

We shall be concerned only with the steady-state solution of these differential equations for which $\mathbf{i} = \mathbf{i}'e^{j\omega t}$ and $\mathbf{V} = \mathbf{V}'e^{j\omega t}$. Thus $j\omega$ may be written for the partial derivative with respect to t and the preceding equations reduce to

$$\frac{d\mathbf{V}'}{dx} = -(R' + jL'\omega)\mathbf{i}'$$

and

$$\frac{d\mathbf{i}'}{dx} = -(G' + jC'\omega)\mathbf{V}' \tag{14.27}$$

for $\mathbf{V}'$ and $\mathbf{i}'$ as functions of x. By differentiating one or the other with respect to x and eliminating one dependent variable, two identical second-order differential equations are obtained for $\mathbf{i}'$ and $\mathbf{V}'$:

$$\frac{d^2\mathbf{i}'}{dx^2} = \gamma^2\mathbf{i}' \qquad \text{and} \qquad \frac{d^2\mathbf{V}'}{dx^2} = \gamma^2\mathbf{V}' \tag{14.28}$$

where γ is written for $\pm[(R' + j\omega L')(G' + j\omega C')]^{1/2}$. This is the familiar second-order differential equation, the solution of which is a periodic function of the spacial variable. Since it is a second-order equation, the general solution contains two constants and it can be seen by substitution that $\mathbf{i}'$ can be written

$$\mathbf{i}' = \mathbf{A}_1e^{\gamma x} + \mathbf{A}_2e^{-\gamma x} \tag{14.29}$$

Substituting in the second of Eqs. (14.26), it is seen that

$$\mathbf{V}' = \mathbf{z}_i(-\mathbf{A}_1e^{\gamma x} + \mathbf{A}_2e^{-\gamma x}) \tag{14.30}$$

where $\mathbf{z}_i$ is written for $[(R' + j\omega L')/(G' + j\omega C')]^{1/2}$. This quantity is analogous to the image impedance and is known as the *characteristic impedance* of the line.

The form of Eqs. (14.29) and (14.30) shows that the current $\mathbf{i}$ and potential $\mathbf{V}$ are progressive waves. $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{z}_i$ are complex constants containing neither x nor t, and the general nature of the solutions is given by the exponential term $e^{j\omega t \pm \gamma x}$. The constant γ is called the *propagation constant* and is analogous to a transfer constant per unit length. It is in general complex, and writing α and β for its real and imaginary parts, the exponential term becomes $e^{\pm \alpha x}e^{j(\omega t \pm \beta x)}$. The first factor is a damping term which represents a decrease in amplitude with the spacial variable x. The second factor represents the progressive wave. For any constant position on the conductors, x is not a variable and the variable factor is $e^{j\omega t}$, which represents a simple sinusoidal variation of the current or potential with the time. At any instant t is not a variable and at various points along the conductors this factor gives the variation with x as $e^{j\beta x}$ which is again a simple sinusoidal variation.

Just as the temporal period τ is $2\pi/\omega$ the spacial period or *wave length* is $\lambda = 2\pi/\beta$. This is the spacial interval between corresponding points of the simple harmonic wave. Consider the particular value of the current or potential that occurs at t and x on the conductors. At a later time, say t', this value of the current or potential will occur at a different position, say x'. If the values of the vector are identical, the values of the exponent are also, *i.e.*, $(\omega t \pm \beta x) = (\omega t' \pm \beta x')$ or $\frac{(x' - x)}{(t' - t)} = \mp \frac{\omega}{\beta}$. The change in position divided by the time interval is the velocity of propagation of the disturbance along the conductors, hence the positive sign in the exponent represents a velocity in the negative direction and the minus sign a velocity in the positive direction, the magnitude of the velocity being ω/β. These results for the progressive wave may be collected and written

$$\tau = \frac{1}{\nu} = \frac{2\pi}{\omega}, \qquad \lambda = \frac{2\pi}{\beta}, \qquad v = \frac{\omega}{\beta} \tag{14.31}$$

The physical phenomena can be more easily visualized by expressing the constants $\mathbf{A}_1$ and $\mathbf{A}_2$ in terms of the terminal conditions. Let $\mathbf{V}'_g$ be the emf. of a generator at the end $x = 0$, $\mathbf{z}_0$ the external impedance at that end, and $\mathbf{i}'_0$ and $\mathbf{V}'_0$ the entering current and potential difference. At this end by Eqs. (14.29) and (14.30)

$$\mathbf{V}'_g - \mathbf{z}_0\mathbf{i}'_0 = \mathbf{V}'_0 = -\mathbf{z}_i(\mathbf{A}_1 - \mathbf{A}_2)$$

or

$$\mathbf{V}'_g = \mathbf{A}_1(\mathbf{z}_0 - \mathbf{z}_i) + \mathbf{A}_2(\mathbf{z}_0 + \mathbf{z}_i)$$

Assume that the far end $x = l$ is terminated by the impedance $\mathbf{z}_l$; then

$$\mathbf{V}'_l = \mathbf{z}_i(\mathbf{A}_2 e^{-\gamma l} - \mathbf{A}_1 e^{\gamma l}) = \mathbf{z}_l \mathbf{i}'_l$$

or

$$0 = \mathbf{A}_1(\mathbf{z}_i + \mathbf{z}_l)e^{\gamma l} - \mathbf{A}_2(\mathbf{z}_i - \mathbf{z}_l)e^{-\gamma l}$$

Solving these equations for $\mathbf{A}_1$ and $\mathbf{A}_2$

$$\mathbf{A}_2 = \frac{\mathbf{V}'_g}{\mathbf{z}_i + \mathbf{z}_0}(1 - \Gamma_0\Gamma_l e^{-2\gamma l})^{-1} \tag{14.32}$$

and

$$\mathbf{A}_1 = \Gamma_l e^{-2\gamma l}\mathbf{A}_2 \tag{14.33}$$

where $\Gamma_0 = (\mathbf{z}_i - \mathbf{z}_0)/(\mathbf{z}_i + \mathbf{z}_0)$ and $\Gamma_l = (\mathbf{z}_i - \mathbf{z}_l)/(\mathbf{z}_i + \mathbf{z}_l)$ are known as the input and output *reflection coefficients*. In terms of these parameters Eqs. (14.29) and (14.30) can be written

$$\mathbf{i}' = \mathbf{A}_2 e^{-\gamma x}(1 + \Gamma_l e^{-2\gamma(l-x)}) \tag{14.34}$$

$$\mathbf{V}' = \mathbf{A}_2\mathbf{z}_i e^{-\gamma x}(1 - \Gamma_l e^{-2\gamma(l-x)}) \tag{14.35}$$

The physical significance of the reflection coefficients is brought out more clearly by expanding the denominator of $\mathbf{A}_2$ by the binomial theorem. Eq. (14.34) then becomes

$$\mathbf{i}' = \frac{\mathbf{V}'_g}{\mathbf{z}_i + \mathbf{z}_0}(e^{-\gamma x} + \Gamma_l e^{-\gamma(2l-x)} + \Gamma_l \Gamma_0 e^{-\gamma(2l+x)} + \Gamma_l^2 \Gamma_0 e^{-\gamma(4l-x)} + \Gamma_l^2 \Gamma_0^2 e^{-\gamma(4l+x)} + \Gamma_l^3 \Gamma_0^2 e^{-\gamma(6l-x)} + \Gamma_l^3 \Gamma_0^3 e^{-\gamma(6l+x)} + \cdots)$$

Each term represents a spacial wave that has traveled a distance equal to the bracket in the exponent. The amplitudes are proportional to the products of powers of the reflection coefficients. Thus the instantaneous value of the current may be thought of as made up of the sum of a large number of current waves that have traveled back and forth on the line suffering partial reflections at the ends. The Γ's represent the fraction of the wave amplitude incident on a terminal that is reflected back into line; the fifth term in the expansion, for instance, represents an amplitude that has been decreased by four reflections, two at each terminal, and has traversed the line four times and in addition the distance x.

By means of the preceding equations, the current in the line and the potential difference between the conductors can be determined at any point and at any time. It is frequently convenient to express these vectors in terms of the input potential $\mathbf{V}'_0$ ($\mathbf{V}'_0 = \mathbf{V}'_g$ if $\mathbf{z}_0 = 0$) using hyperbolic functions

$$\mathbf{i}' = \frac{\mathbf{V}'_0}{\mathbf{z}_i} \frac{\sinh [\gamma(l - x) + \phi]}{\cosh (\gamma l + \phi)}$$

$$\mathbf{V}' = \mathbf{V}'_0 \frac{\cosh [\gamma(l - x) + \phi]}{\cosh (\gamma l + \phi)} \qquad \phi = \tanh^{-1}\left(\frac{\mathbf{z}_i}{\mathbf{z}_l}\right) \tag{14.36}$$

The first of these yields immediately the effective input impedance

$$\frac{\mathbf{V}'_0}{\mathbf{i}'_0} = \mathbf{z}_i \coth (\gamma l + \phi) = \mathbf{z}_i \frac{\mathbf{z}_l + \mathbf{z}_i \tanh \gamma l}{\mathbf{z}_i + \mathbf{z}_l \tanh \gamma l} \tag{14.37}$$

This expression is very useful for calculating the input impedance of a line in terms of its constants and its termination.

For many lines the resistance and conductance per unit length are so small that they may be neglected to a first approximation. In this case γ is a pure imaginary: $\alpha = 0$ and $\beta = \gamma$; $\mathbf{z}_i$ is real and equal to $\sqrt{L'/C'}$. The current and potential are given by Eqs. (14.34) and (14.35) as

$$\mathbf{i} = \mathbf{A}_2 e^{j(\omega t - \beta x)}[1 + \Gamma_l e^{-2j\beta(l-x)}]$$

$$\frac{\mathbf{V}}{\mathbf{z}_i} = \mathbf{A}_2 e^{j(\omega t - \beta x)}[1 - \Gamma_l e^{-2j\beta(l-x)}]$$

The first exponential factor represents the progressive wave character-

istic and the amplitude on a scale of $\mathbf{A}_2$ is given by the second bracket. The vectors are represented graphically in Fig. 14.22. Unity, which is the first term in the bracket, is the vector from 0 to B and the second term is the vector radius BE of magnitude Γ_l making an angle $2\beta(l - x)$ with the real axis. Thus the vectors from the origin to the ends of the diameter represent the current and the potential divided by the characteristic impedance at any instant. On moving along the line the entire diagram rotates about the origin owing to the first exponential factor and the diameter of the circle rotates about B at twice the rate. If the far end of the line is open, $\mathbf{z}_l$ is ∞, and Γ_l is -1, or if the far end is shorted, $\mathbf{z}_l$ is 0 and Γ_l is 1. In either case the circle passes through the origin and the vectors pass through zero periodically. No power is transmitted along the line. Parallel wires forming a line of this type for which both reflection coefficients have unit magnitude are known as *Leçher wires*. They are used as high Q resonant circuits for oscillators and for measuring wave lengths or frequencies in the high radio-frequency range. Consideration of Eq. (14.32) shows that (neglecting line resistance) the denominator vanishes for the resonant frequency and its harmonics. The system is analogous to a vibrating organ pipe, a shorted end is a current antinode and an open end is a potential antinode. At resonance $\beta l = n\pi/2$, where n is an integer and Eqs. (14.36) yield

FIG. 14.22.—Graphical analysis of electromagnetic waves on a resistanceless line.

$$i = \frac{V_0}{Z_i} \sin \omega t \begin{pmatrix} \sin \\ \cos \end{pmatrix} \frac{n\pi x}{2l}$$

$$V = V_0 \cos \omega t \begin{pmatrix} \cos \\ \sin \end{pmatrix} \frac{n\pi x}{2l}$$

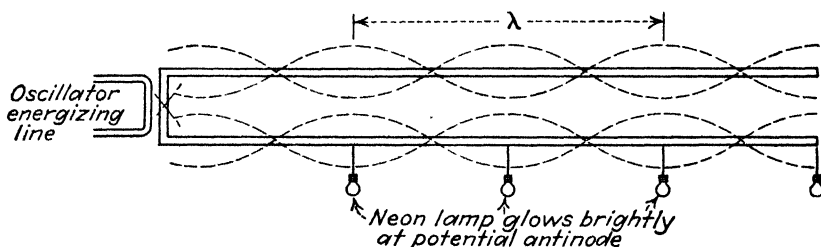



FIG. 14.23.—Lecher-wire system showing standing potential waves.

The choice of the spacial periodic function depends on the terminations, but i and V are always $\pi/2$ out of phase both in space and time. From the resonance condition and Eq. (14.31) $l = n\lambda/4 = nv/4\nu$. From the

total length of the line or from the spacial interval between antinodes, λ can be determined and if v is known, the frequency can be found. It will be seen later that for this type of line v is very closely equal to $1/\sqrt{\kappa_0\mu_0}$ which is 2.998×10^8 m. per second. The interval between current antinodes can be found with a low-resistance ammeter which can be moved along the line, or the interval between potential antinodes can be determined by moving a neon discharge tube along one line. It will glow brightly at potential antinodes owing to capacitative currents. Returning to Fig. 14.22, it is seen that if the far end of the line is terminated in the characteristic impedance, Γ_l vanishes and the circle contracts to the point B. The standing waves disappear and i and V are progressive waves in phase which carry power along the line to the load. If $\mathbf{z}_0$ in Eq. (14.32) is negligible

$$i = V_g\sqrt{\frac{C'}{L'}} \cos(\omega t - \beta x) \qquad \text{and} \qquad V = V_g \cos(\omega t - \beta x)$$

The power $\frac{1}{2}V_g^2\sqrt{C'/L'}$ is completely absorbed by the load at the far end.

The most common type of line is that composed of two parallel copper wires. For this simple geometry the capacity and inductance per unit length are readily calculable if the separation d is large in comparison with the diameter b of the wires. From Eqs. (1.40) and (9.30)

$$C' = \frac{\pi\kappa\kappa_0}{\log_e(2d/b)} \qquad L' = \frac{\mu\mu_0}{\pi}\log_e\frac{2d}{b}$$

If the conductors are large and well insulated so that R' and G' are negligible, the line parameters become

$$\gamma = j\beta = j\omega(L'C')^{\frac{1}{2}} \qquad \text{or} \qquad \beta = \omega\sqrt{\kappa\kappa_0\mu\mu_0}$$

$$\mathbf{z}_i = R_i = \frac{1}{\pi}\sqrt{\frac{\mu\mu_0}{\kappa\kappa_0}}\log_e\frac{2d}{b} \tag{14.38}$$

If κ and μ are approximately unity, $\omega/\beta = v = 1/\sqrt{\kappa_0\mu_0} = 3 \times 10^8$ m. per second. This velocity is independent of the frequency and is the same as the velocity of propagation of electromagnetic radiation in free space (Sec. 16.2). In this case the input impedance may be written

$$R_i = 276 \log_{10}\left(\frac{2d}{b}\right) \text{ ohms}$$

This expression is plotted in Fig. 14.24. Another type of line that is widely used for high-frequency communication work is one composed of coaxial cylindrical tubes. This is known as the coaxial line and has the advantages that the conductor area is large, which reduces the high-frequency resistance, and that the electric and magnetic fields are con-

fined to the annular region between conductors. This eliminates radiation losses and interaction with neighboring lines. The conductors are separated by insulating spacers at intervals along the line. For this case Eqs. (1.23) and (9.31) give the capacity and inductance per unit length

$$C' = \frac{2\pi\kappa\kappa_0}{\log_e (r_1/r_2)} \qquad \text{and} \qquad L' = \frac{\mu\mu_0}{2\pi} \log_e \frac{r_1}{r_2}$$

where r_1 is the inner radius of the outer conductor and r_2 is the outer radius of the inner conductor. These yield the same value for the

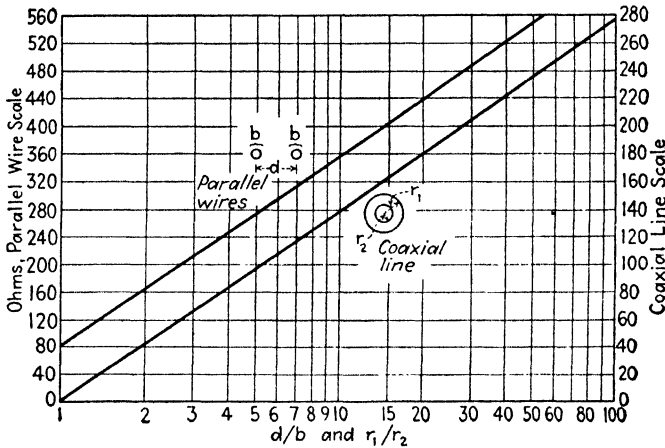



FIG. 14.24.—Characteristic impedance of parallel wires or a coaxial line.

velocity of propagation of the signals and the characteristic impedance is again a pure resistance

$$\mathbf{z}_i = R_i = \frac{1}{2\pi}\sqrt{\frac{\mu\mu_0}{\kappa\kappa_0}} \log_e \frac{r_1}{r_2} \tag{14.39}$$

If the medium between the conductors is air, κ and μ are unity and $R_i = 138 \log_{10} (r_1/r_2)$; this is also plotted in Fig. 14.24.

For the lines discussed in the preceding paragraph R' and G' were entirely neglected, which resulted in a propagation constant that was a pure imaginary and hence there was no attenuation. In practice the resistance and leakage are small, but their effect is not negligible. Assuming that $\omega L' >> R'$ and $\omega C' >> G'$, γ and $\mathbf{z}_i$ may be expanded by the binomial theorem to yield the parameters to a first approximation

$$\begin{aligned} \gamma &= \sqrt{(R' + j\omega L')(G' + j\omega C')} \\ &= j\omega\sqrt{L'C'}\left[1 - j\left(\frac{R'}{\omega L'} + \frac{G'}{\omega C'}\right)\right]^{1/2} \end{aligned}$$

approximately

$$\gamma = \frac{\sqrt{L'C'}}{2}\left(\frac{R'}{L'} + \frac{G'}{C'}\right) + j\omega\sqrt{L'C'}$$

The first factor is the attenuation α and the second the phase constant β. From the latter the velocity of propagation is seen to be the same as for the resistanceless line to this approximation. The attenuation is also independent of the frequency to this approximation which is very desirable for communication circuits. It is proportional to both R' and G' which should therefore be as small as possible. The term containing the conductance is negligible in comparison with the resistance term for well-insulated lines and to this approximation

$$\mathbf{z}_i = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}$$

approximately

$$= \sqrt{\frac{L'}{C'}}\sqrt{1 - \frac{jR'}{\omega L'}} = \sqrt{\frac{L'}{C'}}\left(1 - \frac{jR'}{2\omega L'}\right) \cong \sqrt{\frac{L'}{C'}} = R_i$$

and

$$\alpha = \frac{R'}{2}\sqrt{\frac{C'}{L'}} = \frac{R'}{2R_i} \tag{14.40}$$

Thus the characteristic impedance is approximately a pure resistance independent of the line resistance, the latter affecting only the small reactive component of the impedance. And the attenuation is proportional to R' and inversely proportional to R_i. It is independent of the frequency to this approximation if R' is not a function of ω. This is approximately true at low frequencies, but for the ordinary line of closely spaced small conductors the inductance is not sufficiently greater than the resistance for the approximation to hold. Two methods of increasing the inductance are in general use. One is to surround the conductors by a substance of high permeability which may be accomplished by a serving of permalloy tape. However, this method is expensive and is employed only for submarine cables. The other is to place inductance coils at uniform intervals along the line. This is more economical and is employed generally for land lines, the only disadvantage being the low-pass filter action due to the lumped constants. For large parallel conductors or the coaxial line this loading is in general unnecessary. But when these lines are used for high-frequency transmission, it is necessary to take account of the skin effect. Taking the direct-current resistance per unit length as $1/\pi a^2\sigma$, where a is the conductor radius and σ is the conductivity, the high-frequency approximation of Sec. 13.3 yields

$$R' = \frac{1}{\pi a}\sqrt{\frac{\mu\mu_0\omega}{8\sigma}} = \frac{1}{a}\sqrt{\frac{\nu \times 10^{-7}}{\sigma}}$$

taking μ of the conductor as unity. This expression may be used to calculate the attenuation for either the parallel conductor or coaxial line in the high-frequency range. The attenuation calculated in this way is plotted in Fig. 14.25. Radiation resistance (Sec. 16.5) which increases the attenuation is neglected, as is also the effect of the magnetic

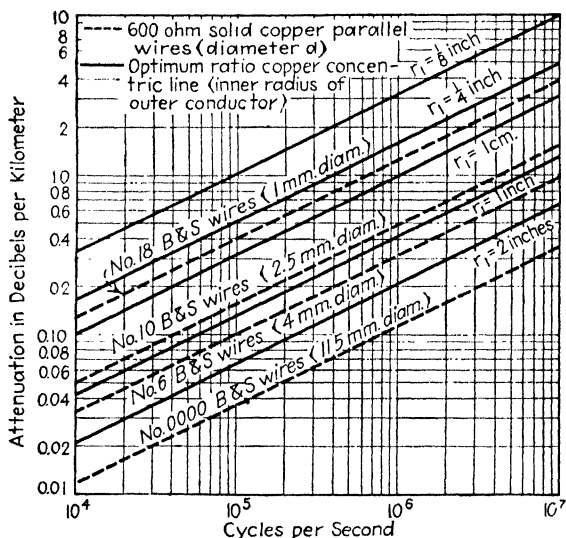



FIG. 14.25.—Attenuation in parallel-wire and coaxial lines.

field of the return wire, which is very small if the separation is greater than 20 times the wire diameter.[1]

Problems

1. A transformer with an efficiency of 90 per cent is used to couple a loud-speaker and series capacity which together present a pure resistance R_l to a vacuum tube with an effective plate resistance R_P. For maximum undistorted output the load presented to the tube should be $2R_P$. Show that the proper turn ratio is

$$\sqrt{R_l/1.8R_P}. \qquad (R_P = R'_s.)$$

2. A line having an input resistance of 600 ohms is to be matched to a tube with a plate resistance of 4,850 ohms by a 90 per cent efficient transformer. Show that the proper turn ratio is 3. $(R_p = R'_s)$.

[1] General references on communication circuits: EVERITT, "Communication Engineering," McGraw-Hill Book Company, Inc., New York, 1937; SHEA, "Transmission Networks and Wave Filters," D. Van Nostrand Company, Inc., New York, 1930; JOHNSON, "Transmission Circuits for Telephonic Communication," D. Van Nostrand Company, Inc., New York, 1929; GUILLEMIN, "Communication Networks," John Wiley & Sons, Inc., New York, 1935.

3. An antenna circuit in the primary of an air-core transformer has a Q_0 of 10 and a natural frequency of 1.5 megacycles. If the coefficient of coupling is 0.2 and the primary and secondary inductances are the same, to what frequency must the secondary be tuned for maximum current at a signal frequency of 1 megacycle?

4. If the secondary circuit of the preceding problem has a Q of 200, what is the critical coupling coefficient?

5. If the secondary circuit of Prob. 3 has a natural frequency of 1.1 megacycles and a Q of 5, what coefficient of coupling will give a maximum secondary current?

6. The *voltage-transfer factor* is the ratio of the potential difference appearing across an element of a circuit to the total potential difference applied to the circuit. Show that the voltage-transfer factor for a condenser in a simple series circuit near resonance is given approximately by

$$F_V = Q_0\left[1 - 2\left(Q_0\frac{\delta\omega}{\omega}\right)^2\right]$$

where $\delta\omega$ is the difference between the applied frequency and the natural frequency of the circuit, if Q_0 is large. Show that the optimum voltage-transfer factor for a condenser in the secondary of an air-core transformer at critical coupling is

$$F_V = \frac{1}{2}\sqrt{\frac{L_2}{L_1}}\sqrt{Q_1Q_2}$$

at resonance.

7. The square root of the ratio of the square of the common reactance between two circuits to the product of the series reactances of like sign in the two circuits separately is known as the coefficient of coupling. Show that the coefficient of coupling in the case of a common capacity C_m is $\sqrt{C_1C_2/C_m^2}$, where C_1 and C_2 are the total series capacities in the two circuits.

8. Show that the coupling between two circuits due to a common inductance L_m can be represented instead by inductive coupling, where

$$M = L_m, \qquad L_1 = L_1' + L_m, \qquad L_2 = L_2' + L_m$$

The primed L's are the inductances in the circuits which are not common.

9. How large an inductance must be placed in series with a 500-cycle generator with an internal resistance of 15 ohms and a load of 100 ohms in order to reduce the third harmonic in the load to 10 per cent of its value as produced by the generator? By how much does this reduce the current of the fundamental frequency? How could the third harmonic be reduced without affecting the current at the fundamental frequency?

10. A coil forms part of a circuit carrying a current of angular frequency ω. A second coil of resistance R and self-inductance L is short-circuited and brought near the first so that the coefficient of mutual inductance is M. Show that the effective resistance presented by the terminals of the first coil to the circuit is increased by

$$\frac{\omega^2M^2R}{R^2 + \omega^2L^2}$$

and that the effective inductance is decreased by

$$\frac{\omega^2M^2L}{R^2 + \omega^2L^2}.$$

11. If the subscript p is used to designate the parameters of the coil in the circuit of the preceding example and the subscript s refers to the parameters of the short-

circuited coil, show that if an effective voltage V_e is applied to the terminals of coil p, the power dissipated in coil s is given by

$$P_s = \frac{V_e^2 R_s M^2\omega^2}{(R_p^2 + \omega^2 L_p^2)R_s^2 + 2M^2\omega^2 R_p R_s + (M^2 - L_p L_s)^2\omega^4 + R_p^2 L_s^2\omega^2}$$

Show that if R_s is variable, the power dissipated in coil s is a maximum when

$$R_s = \omega L_s(1 - k^2)\left(1 + \frac{R_p^2}{\omega^2 L_p^2}\,\frac{1}{(1 - k^2)^2}\right)^{1/2}.$$

(k is the coefficient of coupling between the coils.)

12. Show that if a D'Arsonval galvanometer is placed in series with an alternating emf. the equations governing the motion of the coil are Eqs. (14.17) with

$$\mathbf{z}_{12} = -\mathbf{z}_{21} = -(nabB)$$

in the notation of Sec. (10.5). If there is only resistance in series with the galvanometer, express the current to the instrument as a function of the frequency.

13. From the data given in Sec. (14.3) find the effective elastic coefficient E for the X and Y directions of a quartz crystal if the density of quartz is 2.65 gm. per cubic centimeter.

14. Show that the product of the mid-series and mid-shunt-image impedances of a ladder filter is equal to the product of the series and parallel impedance elements.

15. A low-pass mid-series terminated filter has $L = 1/\pi$ and $C = 1/\pi$. Calculate the cutoff frequency and plot the image impedance as a function of the frequency.

16. A high-pass filter is designed to pass all audio frequencies above 100 cycles and attenuate a 60-cycle hum at at least 70 decibels. How many sections must be used? (Image termination assumed.)

17. In the preceding problem the image impedance is to be 600 ohms at 1,000 cycles. Calculate the values of the inductance and capacity and plot the image impedance as a function of the frequency from 200 to 2,000 cycles.

18. How many sections of a low-pass filter must be used to attenuate the second harmonic of a fundamental frequency about 65 decibels? What is then the attenuation of the third harmonic? (Image termination assumed.)

19. Show that the input impedance of a passive quadripole can be written

$$\mathbf{z}' = \mathbf{z}_{10}\left(\frac{\mathbf{z}_{2s} + \mathbf{z}_l}{\mathbf{z}_{20} + \mathbf{z}_l}\right)$$

20. Show that the ratio of current output to current input for a symmetrical T section is

$$\frac{\mathbf{i}_2}{\mathbf{i}_1} = \frac{\sqrt{\mathbf{z}_{10}(\mathbf{z}_{10} - \mathbf{z}_{1s})}}{\mathbf{z}_{10} + \mathbf{z}_l}$$

21. If the natural frequencies of the series and parallel elements of a simple band-pass filter are the same but the individual reactances are not identical, show that the cutoff angular frequencies are

$$\left(\sqrt{\frac{L_2C_1}{L_1C_2}} + 1\right)^{1/2} \pm \left(\frac{L_2C_1}{L_1C_2}\right)^{1/4}$$

where the subscript 1 refers to the series elements and the subscript 2 to the shunt elements.

22. If the natural frequencies of the series and parallel elements of a simple band-elimination filter are the same but the individual reactances are not identical, show that the cutoff angular frequencies are

$$\frac{1}{4}\left[\left(\sqrt{\frac{L_1C_2}{L_2C_1}} + 16\right)^{1/2} \pm \left(\frac{L_1C_2}{L_2C_1}\right)^{1/4}\right]$$

23. A transmission line of No. 4 bare copper wire ($b = 0.52$ cm.) has a characteristic impedance of 600 ohms; what is its spacing? What is the outer diameter of the inner conductor of a concentric line if the inner diameter of the outer conductor is 5 cm. and it has a characteristic impedance of 60 ohms?

24. A twisted-pair transmission line has an inductance per unit length of 0.5×10^{-6} henry per meter and a capacity per unit length of 6×10^{-11} farad per meter. The power factor is 1.5 per cent independent of the frequency. Calculate the attenuation in decibels per kilometer for the frequencies 0.1, 1, and 10 megacycles. (Neglect leakage conductance.)

25. A transmission line of No. 6 bare copper wires ($b = 0.41$ cm.) 30 cm. apart is three-quarters of a wave length long. Neglecting losses, calculate the vector impedance (*a*) with the end open, (*b*) with the end shorted.

26. Show that the attenuation of a coaxial line for high frequencies is least for a given inner radius of the outer conductor when this radius is approximately 3.6 times the outer radius of the inner conductor.

27. Show that the characteristic impedance is the same as the input impedance of a line that is infinitely long.

28. Show that the attenuation constant of a parallel copper-wire transmission line is

$$1.3 \times 10^{-9}\frac{\sqrt{\nu}}{b \log_{10}(2d/b)} \text{ db per unit length}$$

and of a copper coaxial line is

$$1.3 \times 10^{-9}\frac{\sqrt{\nu}\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}{\log_{10}(r_1/r_2)} \text{ db per unit length}$$

(the conductivity of copper is 5.8×10^7 mho/meter)

29. Show that a T-type section is exactly equivalent to a transmission line of characteristic impedance $\mathbf{z}_i$ and propagation constant γ if the series and shunt elements are given by

$$\mathbf{z}_1 = 2\mathbf{z}_i \tanh \frac{\gamma l}{2}$$

$$\mathbf{z}_2 = \frac{\mathbf{z}_i}{\sinh \gamma l}$$

30. Show that if the constants of a transmission line are related by the equation $L'G' = R'C'$, the attenuation constant per unit length is $\sqrt{G'R'}$, the velocity of propagation of a signal is $1/\sqrt{L'C'}$, and the characteristic impedance is $\sqrt{R'/G'}$. This is known as a distortionless line since the attenuation and velocity are strictly independent of the frequency; the condition implies a larger value of L' than is practically attainable.

31. For the small-gauge unloaded paper-insulated cable widely used in communication work the resistance is large in comparison with the series reactance and the shunt

capacitance is large compared with the leakage conductance. Show that for such a cable

$$\alpha = \beta = \sqrt{\frac{\omega C'R'}{2}} \qquad v = \sqrt{\frac{2\omega}{R'C'}} \qquad z_i = \sqrt{\frac{R'}{\omega C'}}e^{\frac{j\pi}{4}}$$

32. Show that the effective Q value for a resonant line is given by $Q = \beta/2\alpha$ where α is the attenuation constant and β is the phase constant.

33. Show that Eq. (14.10) can be written

$$\left(1 - \frac{\omega_2^2}{\omega^2}\right)\left(1 - \frac{\omega_1^2}{\omega^2}\right) = k^2 - \frac{1}{Q_1^2}\frac{(1 - \omega_2^2/\omega^2)}{(1 - \omega_1^2/\omega^2)}$$

where ω_1 and ω_2 are the resonant angular frequencies of the primary and secondary circuits separately, ω is the applied frequency, k is the coefficient of coupling, and Q_1 refers to the primary. Plot ω_2/ω as a function of ω_1/ω for $Q_1 = 10$ and $k = 0.5$ and 0.1; discuss the significance of these curves and their intersections with the line $\omega_1\omega_2 = \omega^2$.

CHAPTER XV

VACUUM-TUBE CIRCUITS

15.1. Amplifiers with Inductive Loads.—A technical discussion of the problems of radio engineering is beyond the scope of this text, but the vacuum tube is employed so widely as a tool in pure physics and as an element in electrical-engineering circuits that certain of its more important applications will be discussed. It is used as a circuit element to perform three general types of service: (*a*) as a nonlinear element for rectification, harmonic generation, modulation, and demodulation; (*b*) for unidirectional circuit isolation as in the case of the electrostatically screened plate of the tetrode and pentode or in balanced or neutralized circuits employing triodes; (*c*) for the conversion of direct-current power from a battery or generator into alternating-current power throughout the entire frequency range from 0 to about 10^{10} cycles per second as in amplifiers and oscillators. The basic phenomena associated with nonlinear resistances were considered in Chap. V. The use of electrostatic screens for reducing the capacitative interaction between the plate and grid circuits was discussed in Sec. 7.3, and the application of bridge circuits to the balancing or neutralization of vacuum-tube circuits was described in Sec. 13.5, so that it is unnecessary to consider these services specifically in more detail. However, the use of the vacuum tube for the conversion of direct- to alternating-current power in general involves inductive circuits, and as a consequence the earlier discussion of amplifiers in Sec. 7.7 was necessarily limited to voltage amplification in resistance-capacity circuits. The principles involved in the use of a transformer in voltage-amplification circuits are obvious from the discussion in Sec. 14.1 and do not require further consideration.

Voltage amplifiers frequently employ inductive rather than resistive loads since these present a high alternating-current and low direct-current resistance which is an advantage in reducing the requisite direct-current potential for the plate circuit. The circuit is essentially that of the resistance-capacity-coupled amplifier (Fig. 7.24), except that the plate resistance R_l is replaced by an inductance L having a direct-current resistance R. The ratio of the alternating- to direct-current resistance is $(R^2 + \omega^2 L^2)^{1/2}/R$ or $(1 + Q^2)^{1/2}$. From Fig. 7.7 this is seen to be inversely proportional to the difference $E_B - E_b$ for these two plate resistances. (E_B = battery potential, E_b = mean plate potential.) At audio frequencies where iron-core inductances are used it is not possible

to make the Q of the coil very large since it must carry the direct-current component of the plate current which reduces the effective μ of the core; however, a Q of 10 will increase the effective load resistance by approximately 10 without necessitating a larger battery potential. At radio frequencies air-core inductances are used and much higher Q's achieved. A certain frequency distortion is, of course, introduced with this type of load. The potential developed across a load $\mathbf{z}_l$, which is $\mathbf{i}_p\mathbf{z}_l$, is from Eq. (7.10)

$$\mathbf{e}_l = \frac{\mu_p \mathbf{z}_l}{r_p + \mathbf{z}_l}\mathbf{e}_g$$

If the Q of the inductance is considerably greater than unity, the amplification factor $A = e_l/e_g$ is approximately $\mu_p\omega L/(r_p^2 + \omega^2L^2)^{1/2}$. This rapidly approaches μ_p as the product ωL increases. For frequencies (ν's) greater than $r_p/2L$, A is within 5 per cent of μ_p; hence this type of distortion is appreciable only at low frequencies. If the inductance is parallelled by a condenser to form a tuned load, a still higher effective resistance can be achieved. The resonant resistance with a perfect condenser is RQ^2 (Sec. 13.4), but since the frequency discrimination is high, this type of load is suitable only for the amplification of a very narrow frequency band.

15.2. Class A Power Amplifiers.—The general principles of the linear amplification of an alternating current by means of a vacuum tube were discussed in Sec. 7.2. However, the assumption of linearity is at best an approximation, and where the accurate reproduction of a wave form is of importance, it is necessary to consider the distortion as well as the power output. Frequency discrimination or distortion and delay or phase distortion are largely matters of the external circuit and have been discussed in preceding sections. The distortion that enters owing to the nonlinearity of the tube characteristics is known as amplitude distortion. The class A amplifier is by definition one in which plate current flows at all times and the grid excursion is sufficiently small so that the linear approximation of Sec. 7.2 is applicable. However, the vacuum tube is essentially a nonlinear device and for any finite alternating-current grid potential e_g there will be some distortion of the plate-circuit vectors. Even-order harmonic distortion can be reduced by the use of two tubes in a balanced circuit or by the use of a pentode instead of a triode. In what follows the power output will be considered in relation to the distortion that is involved.

The fundamental triode circuit is illustrated in Fig. 15.1. A transformer is shown in the plate circuit for purposes of generality since some impedance-matching device is usually necessary between the tube and the load. The plate potential is applied through the inductance L

which provides a low resistance direct-current path. The condenser C has a low impedance for the alternating-current components to be amplified and the combination prevents the direct-current component of the plate current from flowing through the transformer primary and reducing the effective permeability of the core. The effective resistance presented to the tube is approximately $R_l' = R_l/n^2$. In general the

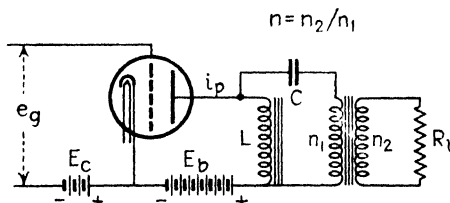



FIG. 15.1.—Single-tube amplifier and output transformer.

load is not a pure resistance and by Eq. (7.10) the power consumption is

$$P = i_p^2 R_l' = \frac{\mu_p^2 e_g^2}{(r_p + R_l')^2 + X_l'^2} R_l' = \frac{\mu_p^2 e_g^2 n^2 R_l}{(n^2 r_p + R_l)^2 + X_l^2}$$

Or, if the load is a pure resistance, which is the optimum case, the average power is given by

$$P = \frac{\mu_p^2 E_{gm}^2 R_l'}{2(r_p + R_l')^2} \tag{15.1}$$

where E_{gm} is the peak or maximum value of the alternating-current voltage applied to the grid. This expression is based upon the linear approximation and implies very small values of E_{gm}. If it applies, the maximum power transfer occurs for $r_p = R_l'$, but for appreciable power to be obtained E_{gm} cannot be restricted to very small values. In consequence the linear approximation is really inadequate and graphical methods must be employed.

It is evident from Fig. 7.6 that the curvature of the characteristic is most pronounced in the region of small plate currents. Hence the plate current must not be allowed to fall to too low a value if distortion is to be avoided. Let us say that the smallest permissible value is $I_{b\,\min}$. Likewise the maximum positive excursion of the grid is determined by the point at which distortion sets in owing to the flow of grid current. This occurs when the potential of the grid reaches approximately that of the cathode, *i.e.*, $e_c = 0$. The third limiting factor is that the product $E_b I_b$ is determined approximately by the allowable plate dissipation of the tube. Hence the region in which the operating point can lie is the shaded portion of Fig. 15.2 which is bounded by the line $I_{b\,\min} = \text{const.}$, the curve $e_c = 0$, and the hyperbola $E_b I_b = \text{const.}$ For a given operating

potential, E_b, the optimum direct-current plate current, I_b, can be shown to be

$$I_b = \frac{(I' + 3I_{b\,\mathrm{min}})}{4}$$

where I' is the ordinate of the curve $e_c = 0$ at the abscissa E_b. For from Fig. 15.2 the power output is given by

$$P = \tfrac{1}{2}\, OB\, DB = \tfrac{1}{2} r_p\, OB\, BC$$

where r_p is the reciprocal of the slope of the characteristic $e_c = 0$ which is assumed straight. OB and BC both depend on the choice of I_b and

FIG. 15.2.—Triode power output and distortion.

the load resistance but the distance AC is a constant and for approximately linear operation AO must be approximately equal to OB. Thus $AC = AO + OB + BC = 2OB + BC$, and hence

$$P = \tfrac{1}{2} r_p (AC - 2OB) OB$$

The condition for maximum power is obtained by setting $dP/d(OB) = 0$ which yields $OB = AC/4 = BC/2$. Since $I_b = I_{b\,\mathrm{min}} + OA$ and

$$I' = I_{b\,\mathrm{min}} + AC$$

the above condition for I_b is obtained. Since $r_p = DB/BC$, the proper value for the load resistance R_l' is given by

$$R_l' = \frac{DB}{OB} = r_p \frac{BC}{OB} = 2r_p$$

The power output increases with battery potential and reaches its maximum value when the operating point is at O_{max} {the intersection of the curves $E_b I_b =$ const. and $i_b = [I'(e_b) + 3I_{b\,\mathrm{min}}]/4$}. If $R_l' = 2r_p$

$$P = \frac{\mu^2 E_{gm}^2}{9r_p} \tag{15.2}$$

This equation may be used to obtain the approximate output in terms of the plate resistance and the maximum signal applied to the grid. Since E_b/μ is of the order of the cutoff grid potential, the maximum permissible value of E_{gm} is of the order of $E_b/2\mu$ and from Eq. (15.2) the maximum power subject to the limitations imposed by distortion is of the order of $E_b^2/36r_p$. Likewise the maximum plate-circuit efficiency, which is the ratio of this quantity to E_bI_b is approximately $E_b/36r_pI_b$. This is quite small. Inserting the constants of a type 10 tube, for instance,

$$(E_b = 425 \text{ volts}, I_b = 0.018 \text{ amp.}, r_p = 5{,}000 \text{ ohms})$$

yields $P = 1$ watt and a plate-circuit efficiency of about 16 per cent. It must be borne in mind that these are extreme values and for good reproduction it may be necessary to use values of E_{gm} that are smaller than $E_b/2\mu$.

The power output per tube can be increased and the amplitude distortion greatly reduced by the use of the balanced or push-pull amplifier shown in Fig. 15.3. To a good approximation the dynamic characteristic of a tube is parabolic over a limited range, *i.e.*, $i_b = I_b + ae_g + be_g^2$. This leads to second-harmonic distortion for a single tube, and it was seen in Sec. 7.4 that even harmonics are removed by a balanced circuit.

FIG. 15.3.—Push-pull amplifier.

The class *A* operation of a balanced circuit approaches more nearly the ideal of linearity than any other vacuum-tube circuit. The input transformer applies equal and opposite e_g's to the two tubes and the plate circuits are coupled together by the center-tapped output transformer. Thus $e_{p1} = -e_{p2}$ and the circuit can be handled graphically by reversing the voltage scale of the plate characteristics of tube 2 and plotting them beneath the characteristics of tube 1 in such a way that the mean operating plate potential E_b coincides on the two scales. This is shown in Fig. 15.4, using representative data for a type 45 tube. The composite characteristics (primary current in the output transformer as a function of the plate potential of either tube) are obtained by adding the characteristics for equal and opposite alternating-current grid potentials (e_g's) and are seen to be a set of approximately equally spaced straight lines. Thus the operation is closely linear for quite large values of E_{gm}. From the method of formation of the composite characteristics their slope is approximately twice that of the characteristic of a single tube in the neighborhood of the quiescent point. Furthermore, the operation is so closely linear that the maximum power output occurs for

$$R_l' = r_{p\,\text{eff}} = \frac{r_p}{2}$$

Using this value of the load and dividing r_p by 2, Eq. (15.1) yields $P = \mu_p^2 E_{gm}^2/4R_p$. On comparing this with Eq. (15.2) the push-pull circuit is seen to produce more than twice the power available from a single tube. In practice the factor is larger than this simple analysis would indicate as considerably larger values of E_{gm} can be used than in the case of a single tube without undue distortion. The push-pull

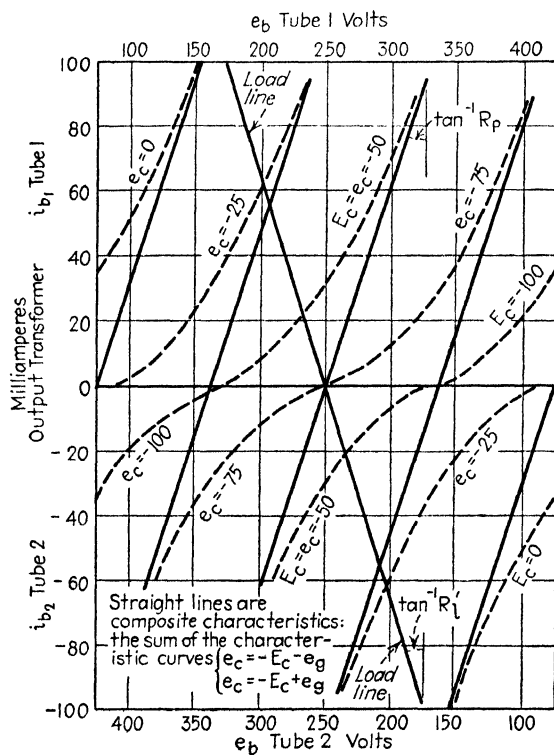



FIG. 15.4.—Composite characteristics of the two tubes in a push-pull circuit (type 45).

circuit has further advantages in that emfs. introduced in the common battery circuit have no effect on the output to a first approximation. Thus, if the plate potential supply is obtained from a rectifier, it need not be so well filtered as for a single tube. Likewise the method of obtaining the grid biasing potential by means of a resistance in the cathode-plate circuit by-passed by a condenser has less degenerative effect than for a single tube. This simple analysis has assumed identical tubes and the closer this ideal is approached, the more satisfactory is the circuit. If well-matched tubes are chosen, slight inequalities may be approximately compensated by individual adjustment of the grid-biasing potentials.

Finally, the pentode is frequently used as a class *A* amplifier for, as the amplification constant is generally larger for a pentode than for a triode, a larger power output can be obtained for the same grid excitation, E_{gm}. The circuit is essentially that of Fig. 15.1 for the potentials of the additional electrodes are constant. The screen grid is maintained at the direct-current potential specified for the tube (alternating-current components of the current to it are by-passed to ground by a condenser), and the suppressor grid is generally connected

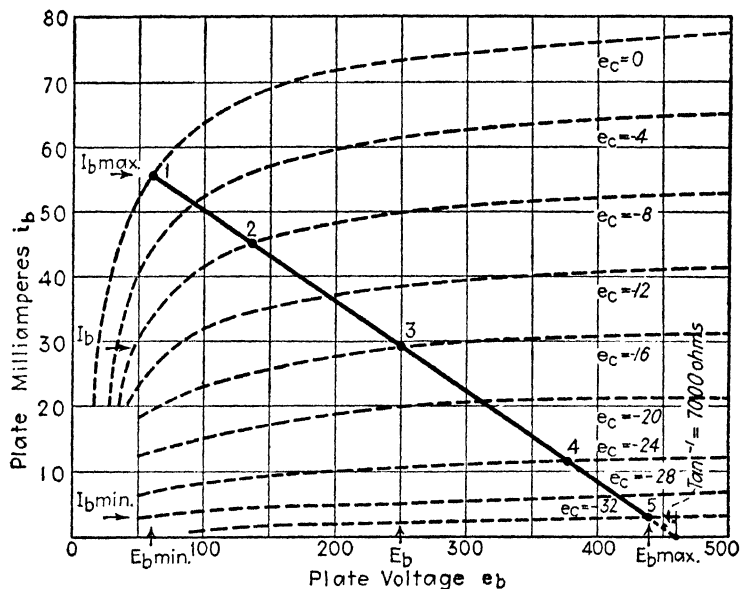



FIG. 15.5.—Power-pentode characteristics (type 47).

directly to the cathode. A representative family of characteristics is shown in Fig. 15.5 (type 47 tube). In distinction to the triode the load resistance for small distortion should be considerably smaller than the plate resistance. E_b is fixed by the allowable plate dissipation and the maximum current by the characteristic $e_c = 0$. A consideration of the load line shows that the same type of distortion appears at both extremes of e_g as the load resistance is increased. This property permits the elimination of second-harmonic distortion by the proper choice of load resistance, and in consequence there is less advantage in the push-pull operation of pentodes. However, the third harmonic is in general considerably larger than in balanced circuits employing triodes. The power output and amplitude distortion can be obtained from the static characteristics just as in the case of triodes.

The amplitude distortion, *i.e.*, the amplitudes of the harmonics present, can be obtained from the dynamic characteristic by the graphical methods of Chap. V and

a Fourier analysis of the resulting wave. Since the characteristic is retraced each cycle, the wave is symmetrical about either extreme on the characteristic and can always be written

$$i = I_0 + I_1 \cos \omega t + I_2 \cos 2\omega t + I_3 \cos 3\omega t + I_4 \cos 4\omega t + \cdots \tag{15.3}$$

To obtain the amplitude of the constant term and harmonics up to the nth, $(n + 1)$ ordinates must be measured and the $(n + 1)$ simultaneous equations solved. (t or ωt is chosen as 0 at an ordinate corresponding to an extreme excursion along the characteristic.) A knowledge of the amplitude of the first three harmonics is adequate for many purposes and these may be obtained by evaluating i at ωt intervals of $\pi/3$. This particular choice also corresponds to evaluation at equal intervals of e_g. The analysis may also be performed very simply from the family of static characteristics and the load line. For instance, the ordinates corresponding to the points 1, 2, 4, and 5 of Fig. 15.5 are all that are necessary to determine the harmonic amplitudes up to the third. Designating these ordinates by the corresponding subscripts, Eq. (15.3) becomes at these points

$$\begin{array}{lll} \omega t = 0, & e_g = E_{gm} & i_1 = I_0 + I_1 + I_2 + I_3 \\ \omega t = \frac{\pi}{3}, & e_g = \frac{E_{gm}}{2} & i_2 = I_0 + \frac{1}{2}I_1 - \frac{1}{2}I_2 - I_3 \\ \omega t = \frac{2\pi}{3}, & e_g = -\frac{E_{gm}}{2} & i_4 = I_0 - \frac{1}{2}I_1 - \frac{1}{2}I_2 + I_3 \\ \omega t = \pi, & e_g = -E_{gm} & i_5 = I_0 - I_1 + I_2 - I_3 \end{array}$$

Solving for the harmonic amplitudes

$$\begin{array}{ll} I_0 = \frac{1}{3}[(i_2 + i_4) + \frac{1}{2}(i_1 + i_5)] & I_1 = \frac{1}{3}[(i_1 - i_5) + (i_2 - i_4)] \\ I_2 = \frac{1}{3}[(i_1 + i_5) - (i_2 + i_4)] & I_3 = \frac{1}{6}[(i_1 - i_5) - 2(i_2 - i_4)] \end{array}$$

Even harmonics are dependent on sum terms and odd harmonics on difference terms. If the points 1 and 2, and 4 and 5 are symmetrical about 3, $I_0 = i_3$ and the direct-current component is independent of the signal, *i.e.*, the signal is not rectified. In the case of pentode characteristics the operating line may be so chosen that

$$(i_1 + i_5) = (i_2 + i_4)$$

and the second harmonic vanishes; this is not practicable in the case of triodes. If R_l is the load resistance, the direct-current power dissipated therein is $I_0^2R_l$. The power at the fundamental frequency and its harmonics is given by $\frac{1}{2}I_1^2R_l$, $\frac{1}{2}I_2^2R_l$, etc.

15.3. Class B Amplifiers.—The output of an amplifier operated over only its linear range is so limited and its efficiency is so poor that it is inadequate for most power purposes. The class B amplifier is one for which the quiescent grid potential (fixed negative bias) is such that when there is no signal ($e_g = 0$), a negligible plate current flows. This greatly improves the efficiency since it reduces the large direct-current losses associated with the class A amplifier. Plate current flows for positive-grid excursions but not for negative ones. Thus the positive portion of the e_g cycle appears in the output, but the negative portion is suppressed, much as in the case of a rectifier. The graphical treatment of a single tube by means of the dynamic characteristic is illustrated in Fig. 15.7. The plate current is obviously badly distorted; in fact, if

the dynamic characteristic is considered as approximately a straight line, the plate-current loops are half sine waves for which the Fourier analysis is given in Sec. 5.9. However, if suitable precautions are observed, it is possible to operate the push-pull circuit class B, the second tube inserting the portion of the cycle rejected by the first. The circuit is somewhat analogous to the full-wave rectifier, except that the output transformer reverses the sense of alternate loops so that the wave is reproduced with approximate correctness. The composite dynamic characteristic is indicated in Fig. 15.6. The grid excursion is well beyond the class A region in which the tubes compensate one another, but the balanced circuit eliminates the second harmonic to a first approximation, and if the curvature of the characteristic is not too great, the third harmonic is small. Thus this circuit may be used for the accurate reproduction of wave forms and with a considerable power output. The plate supply must be capable of handling the power requirements of this circuit and as the fluctuation of i_b is large, the power supply should have a low internal resistance. This large fluctuation of the plate current renders self-bias (bias produced by the iR drop in a resistance in the cathode circuit) impracticable, and the grid bias, if any, must be supplied by a battery or generator. As a consequence tubes having a high amplification constant which can be operated at zero bias have been designed for this purpose. These are not as efficient as if a biasing potential were used, but for many purposes they are much more convenient. A further restriction is imposed on the circuit by the flow of grid current for the extreme values of e_g. In order to avoid distortion under these circumstances the external resistance of the grid circuit must be as small as possible. In consequence the primary circuit of the input transformer, which generally consists of the plate circuit of the preceding tube, should have a low resistance and be capable of supplying adequate power for the grid losses. Frequently the input transformer is of the stepdown type ($m_1 > m_2$) to reduce the effective resistance presented to the grid circuit.

FIG. 15.6.—Push-pull class B operation.

If the dynamic characteristic is approximately linear, a single tube is a satisfactory amplifier for a wave that is modulated even up to 100 per cent. This is illustrated graphically in the lower portion of Fig. 15.7. A simple modulated wave may be written (Sec. 5.6)

$$V = V_1 \cos \omega_1 t(1 + a \cos \omega_2 t)$$

where $a = V_2/V_1$, is the ratio of the low-frequency modulation to the amplitude of the unmodulated high-frequency wave. Though the negative-grid excursions are not reproduced in the plate circuit, the low-frequency character of the modulated envelope is not lost. The non-linearity of the characteristic may distort this envelope, but this is not worse than ordinary class A distortion unless the characteristic has a very pronounced curvature. The push-pull circuit may, of course, be employed to reduce even-harmonic distortion in the amplification of a modulated wave. The amplifier may be even biassed beyond cutoff (class C) and a modulated envelope will be reproduced, provided the

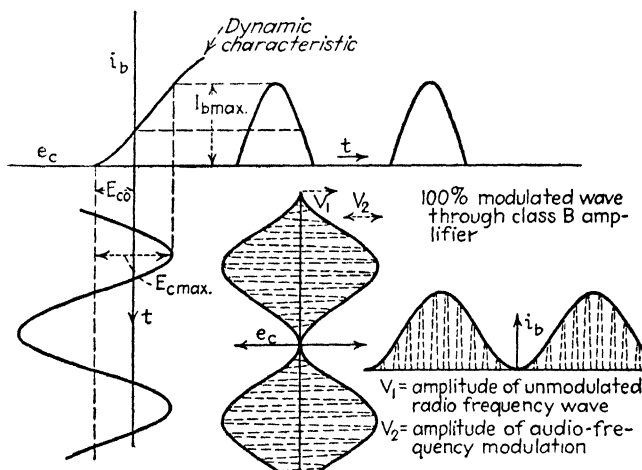



FIG. 15.7.—Graphical analysis of the class B amplifier.

bias beyond cutoff is not greater than $V_1 - V_2$, as may be seen from an inspection of the graphical analysis.

It is very difficult to calculate the actual power output, efficiency, and harmonic content of a nonlinear amplifier and the adjustments are generally empirical. However, a certain insight is gained into its operation by considering a partially linear analysis.[1] Assume that the dynamic characteristic of the tube is linear to a first approximation so that at least part of the positive excursion of the grid results in a portion of a sine wave of plate current. This in general introduces two sources of error: (a) distortion due to the curvature near cutoff, which, however, is usually not serious; (b) distortion due to the curvature of the characteristic for large values of e_g, which is not serious for class B amplifiers as this region is not invaded. Carrying the linear assumption farther Eqs. (7.6) and (7.8) are assumed to hold in the conducting region and they yield

[1] EVERITT, *Proc. I.R.E.*, **22**, 152 (1934).

$$i_p = s_p\left(e_g + \frac{e_p}{\mu_p}\right) = s_p e'_g \tag{15.4}$$

Figure 15.8 is a plot of the plate current on this primed potential scale for which the applied grid potential is supplemented by the plate potential divided by the amplification constant. On this scale the magnitude of the negative grid bias, E'_c, is given by

$$E'_c = E_c - E_0 = E_c - \frac{E_b}{\mu_p} \tag{15.5}$$

And on this scale the magnitude of the sinusoidal grid potential is the applied excitation, E_g, plus the fluctuating component of the plate poten-

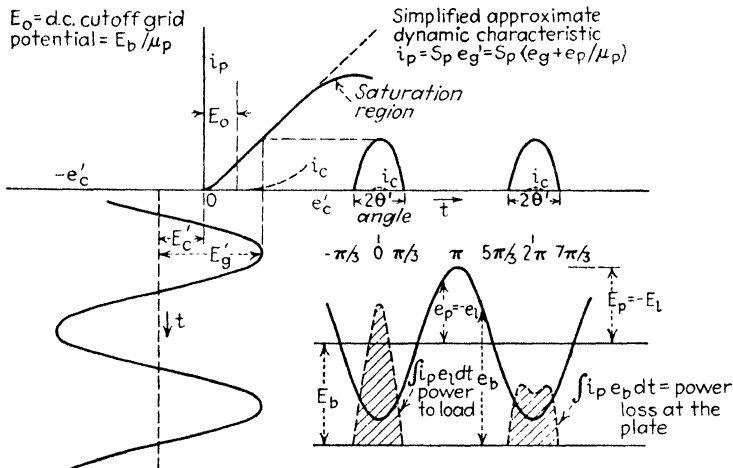



FIG. 15.8.—Approximate analysis of a class B or class C amplifier.

tial, divided by μ_p. Assume that the load is a parallel circuit tuned to the fundamental which presents a negligible resistance to the harmonics. Then $R_l I_{p1} = E_l = -E_p$ and

$$E'_g = E_g - \frac{R_l I_{p1}}{\mu_p} \tag{15.6}$$

where I_{p1} is the amplitude of the fundamental component of the plate current. From Fig. 15.8 the plate current is the portion of the sinusoidal curve induced by E'_g which lies above the zero axis, or

$$i_p = s_p(E'_g \cos \omega t - E'_c) \tag{15.7}$$

This wave contains many harmonics, but since the plate load is tuned to the fundamental, all the potential components across it except E_{p1} are in general negligible. Hence only the fundamental power and average (direct-current) plate dissipation need be calculated.

The constant component of the plate current, I_b, and the amplitude of the fundamental component, I_{p1}, can be obtained by expanding Eq. (15.7) in a Fourier series. From Appendix *B* and Fig. 15.8

$$I_b = \frac{1}{2\pi}\int_{-\pi}^{\pi} i_p \, d(\omega t) = \frac{1}{\pi} s_p \int_0^{\theta'} (E'_g \cos \theta - E'_c) d\theta$$

where θ' is the cutoff angle. From the construction of Fig. 15.8

$$\cos \theta' = \frac{E'_c}{E'_g} \tag{15.8}$$

hence the result of the integration can be written

$$I_b = \frac{E'_g s_p}{\pi}(\sin \theta' - \theta' \cos \theta') \tag{15.9}$$

Similarly the amplitude of the fundamental component is given by

$$\begin{aligned} I_{p1} &= \frac{1}{\pi}\int_{-\pi}^{\pi} i_p \cos \omega t \, d(\omega t) = \frac{2}{\pi} s_p \int_0^{\theta'} (E'_g \cos \theta - E'_c) \cos \theta \, d\theta \\ &= \frac{E'_g s_p}{\pi}\left(\theta' - \frac{1}{2} \sin 2\theta'\right) \end{aligned} \tag{15.10}$$

In the special case of class *B* operation which is the one under immediate consideration $\theta' = \pi/2$ and

$$I_{p1} = \frac{E'_g s_p}{2} = \frac{\mu_p}{2r_p}\left(E_g - \frac{R_l I_{p1}}{\mu_p}\right)$$

or

$$I_{p1} = \frac{\mu_p E_g}{2r_p + R_l}$$

Thus the effective plate resistance is twice the plate resistance of the tube. Similarly, comparing Eqs. (15.9) and (15.10) at $\theta' = \pi/2$

$$I_b = I_{p1}\frac{2}{\pi}$$

Also, the plate-circuit efficiency is defined as the ratio of the output power to the power supplied to the plate circuit.

$$\begin{aligned} \text{Plate efficiency} &= \frac{E_{p1} I_{p1}}{2E_b I_b} \\ &= \frac{\pi}{4}\frac{E_{p1}}{E_b} \end{aligned} \tag{15.11}$$

Since the limiting value of the fluctuating plate potential is E_b itself the maximum efficiency is $\pi/4$ or about 78.5 per cent. Actually E_{p1} is

less than E_b and the efficiency is generally of the order of 60 per cent. This efficiency cannot be achieved in the amplification of a modulated wave. For from inspection of the lower portion of Fig. 15.7 it is seen that the amplitude of the carrier frequency is only about one-half of the peak amplitude of the wave, which reduces the actual efficiency by a factor of approximately four.

15.4. Class C Amplifiers.—A class *C* amplifier is one in which the negative-grid bias is so great that plate current flows for much less than half of a cycle. The preceding analysis is only very approximately applicable to this case, since for a small portion of the positive half cycle the grid becomes sufficiently positive to invade the nonlinear region of the dynamic characteristic and the plate-current loops cease to be portions of sine waves. Grid current flows during almost as large a portion of the cycle as does plate current; in fact, the grid bias for this type of amplifier can be obtained largely from the iR drop in a resistance in series with the grid. This is known as a *grid leak,* the alternating-current component of the grid potential being by-passed across it by means of a condenser. Only a fraction of the grid bias should be so obtained, however, since in the absence of excitation the grid potential would be insufficiently negative to limit the plate dissipation to allowable values. The load for this type of amplifier is always a tuned circuit and it may be considered that the circulating current in this circuit is kept in oscillation by the periodic impulses that are supplied by the plate generator when the conductance of the tube rises on positive grid swings. The circulating power at the fundamental frequency in the tuned plate circuit (frequently known as the *tank circuit*), which is $E_{p1}I_{p1}/2$, is proportional to the rate at which the circuit receives power from the plate generator. This is affected to a certain extent by the grid excitation even beyond the linear portion of the dynamic characteristic. But in general with this type of amplifier the grid is driven positive well into the saturation region; the output thus ceases to be proportional to E_g and the amplifier cannot be used for the amplification of a modulated wave.

The adjustments of a class *C* amplifier are largely empirical and a discussion of them is beyond the scope of this treatment; however, a brief graphical analysis will indicate the dependence of output and efficiency on the circuit parameters. In the lower portion of Fig. 15.8 the sinusoidal plate potential due to the generator E_B and the potential across the load is plotted in the proper relative phase to the plate current. The maximum value of the plate or load potential, E_{p1} or E_l, is limited by E_B, and in general the grid is driven sufficiently positive so that the difference between E_l and E_B is small. Since the circulating power is proportional to E_l^2, the output of this type of amplifier is approximately proportional to E_B^2. If this proportionality is accurately maintained, the amplifier

may be linearly modulated by varying the effective potential applied to the plate circuit. The energy dissipated at the plate per cycle is $\int i_p e_b \, dt$ over the interval of plate-current flow. This is indicated by the shaded region to the right in the figure. The energy supplied to the load per cycle is $\int i_p e_l \, dt$ over the interval of plate-current flow and is indicated by the shaded region to the left in the figure. As the grid bias is increased to decrease the interval of flow, θ', the plate dissipation is reduced. Since the difference between E_B and E_P is approximately constant, an increase in E_B has little effect on the plate dissipation, but the power supplied to the load increases rapidly. Therefore the efficiency increases with E_B and theoretically approaches 100 per cent. However, the plate-current loops cease to be portions of sine waves for large positive-grid excursions (see Fig. 15.19), owing to the curvature of the characteristic which is largely due to cathode saturation. This generally limits the fundamental efficiency to 70 or 80 per cent. A high-Q series circuit, resonant at the fundamental, in series with the plate and load, may be used to reduce harmonic components of the plate current and increase the efficiency. (The direct-current component of the plate current is by-passed across this combination by means of an inductance.) The grid loss is in general quite large (of the order of a tenth the plate loss) and must, of course, be supplied by the input circuit. The cathode-heating power is also appreciable in comparison with these. All losses must be taken into account for over-all efficiency and for calculating the total heat energy that must be removed from the tube.

15.5. Feedback Amplifiers.[1]—The behavior of the amplifiers discussed in the preceding sections is conditioned very largely by the vacuum tubes that compose the essential part of the circuit. It is not necessary that this should be the case, and in fact in the instance of the cathode-follower amplifier of Sec. 7.7 it was seen that in limiting circumstances the amplification became unity and the output resistance became very small for a large amplification constant. This is an illustration of the general group of circuits known as *feedback amplifiers*. These are characterized by one or more interconnections whereby the amplified signal influences the signal being amplified. This is an effect which, of course, is to some extent inherent in the vacuum tube itself, for the potential of the grid is determined in part by that of the plate through the interelectrode capacity. However, such interaction is minimized by screening grids, and in the feedback amplifier the phase and amplitude of the interaction

[1] General amplifier references: REICH, "Theory and Applications of Electron Tubes," McGraw-Hill Book Company, Inc., New York, 1939; TERMAN, "Radio Engineering," McGraw-Hill Book Company, Inc., New York, 1947; BODE, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., 1945; VALLEY and WALLMAN, "Vacuum Tube Amplifiers," McGraw-Hill Book Company, Inc., New York, 1948.

are controlled by the external circuit connections. In this way the performance of the amplifier can be made to depend primarily on the characteristics of this external circuit, and numerous advantages may clearly result through the opportunity of choosing this circuit in accordance with the characteristics desired for the amplifier. One of the most important of these advantages is that the linearity of the amplifier can be greatly improved. By choosing linear circuit elements for the feedback circuit the inherently nonlinear attributes of the vacuum tube itself can be reduced to negligible consequence. Also a wide range of frequency characteristics can be chosen for this circuit and the frequency characteristics of the amplifier thus determined.

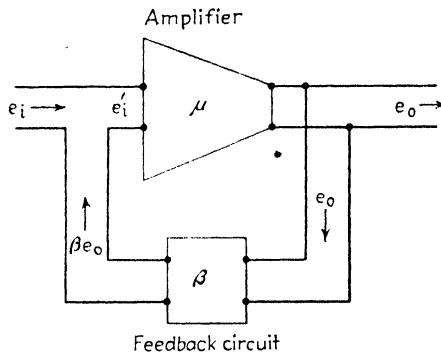



FIG. 15.9.—Schematic feedback amplifier.

Figure 15.9 is a simple block diagram of a feedback amplifier or of a feedback stage in a more complicated amplifier circuit, illustrating the general principle of operation. The amplifier box is drawn with sloping sides to call attention to its unidirectional nature. The amplification of the amplifier unit is taken to be μ, and β represents the complex transfer constant of the feedback circuit in the direction of the arrows. From the diagram it is clear that the relations between the signal voltages shown are given by

$$e_i' = e_i + \beta e_0 \qquad \text{and} \qquad e_0 = \mu e_i'$$

or eliminating e_i' the ratio of e_0 to e_i, the effective amplification constant of the feedback amplifier is

$$\frac{e_0}{e_i} = \frac{1}{\beta}\frac{\mu\beta}{1 - \mu\beta} \tag{15.12}$$

The quantity β is generally complex, and Eq. (15.12) gives both the phase and amplitude of the output in terms of the input. Certain interesting cases are distinguished by the value of the product $\mu\beta$. If this product is very large, the effective amplification becomes $-1/\beta$, and the amplifi-

cation is determined entirely by the characteristics of the feedback circuit. This has the great advantage of eliminating the variability in amplification inherent in μ. If μ is large, the approximation may be adequate even for small β, and hence considerable amplification may still be achieved. Also it is seen that the amplification is independent of the load impedance unless the load is part of the feedback network, in which case the amplification can be made to vary in a desired way with load impedance. If the product $\mu\beta$ is not large, e_0/e_i depends critically on its value, which is in general, of course, a function of the frequency through β. The nonlinearity is enhanced rather than minimized, and the practical applications are of an entirely different nature. If $\mu\beta$ is real, the effective amplification may be either less than or greater than μ; in the former case the amplifier is said to be *degenerative*, and in the latter case *regenerative*. If the product is real and exceeds 1, the characteristic is negative and oscillations may be induced in the associated circuits (Secs. 5.6 and 15.6). The general criterion for stability, which is due to Nyquist, is that the path of $\mu\beta$ in the complex plane as a function of the frequency shall not encircle the point $(1,0j)$. If this criterion is not fulfilled, oscillations will occur at some frequency.

Amplification Limits.—The limit on useful amplification is not imposed by instability but rather by unwanted electrical disturbances in the circuit which are amplified along with the signal and may mask it completely if they are present at a higher level. These disturbances are known generically as *noise*, and the signal-to-noise ratio determines the amplifier's optimum performance. Certain sources of noise are avoidable by the choice of circuit elements and careful design and construction. These include erratic resistors, imperfect contacts, and mechanical vibrations. Other noise sources such as fluctuating cathode emission, ionization in residual gas, and secondary emission from the elements can generally be minimized and need not limit ultimate amplification. A basic limit, however, is imposed by fluctuations that are inherently of thermal origin. These can be reduced by lowering the temperature of critical elements of the circuit but can never be completely eliminated. They arise from the statistical variation of emission of electrons from a cathode (Sec. 6.5) or the statistical variation of electron motion in circuit elements (Sec. 6.3). It can be shown[1] that the mean square noise voltage in a frequency range f appearing across a resistance R at a temperature T if $4RkTf$, where k is Boltzmann's constant. This is known as *Johnson noise* and is frequently the limiting factor when large input resistances are used. In the case of a reactive circuit the relation between the mean square values of the current through L and potential across C is given by $L\overline{i^2} = C\overline{V^2} = kT$; thus noise cannot be eliminated by reducing R. Also the fluctuation in electron emission from a cathode gives rise to an effective mean square potential fluctuation of the grid that can be written $2e\bar{\imath}f/s_g^2$, where e is the electron charge, $\bar{\imath}$ is the mean emission current, f is the frequency interval, and s_g is the trans-conductance of the tube. This is generally known as *shot* noise and by comparison with the expression for Johnson noise can be written in terms of an effective input resistance at a given temperature and emission current. The value of one or another of these noise voltages at the point of minimum signal in an amplifier circuit determines whether or not the signal can be

[1] BARNES and SILVERMAN, *Rev. Mod. Phys.*, **6**, 162 (1934).

detected upon subsequent amplification. No purpose is served in increasing the amount of amplification beyond the point at which input noise is evident at the output.

15.6. Characteristics with Negative Slopes and Instability.—The characteristic curve of a two-terminal device is the graphical relation between the current flowing through the terminals and the potential difference across them, the latter being taken as positive when it is in such a direction as to oppose the flow of current. The product of abscissa and ordinate of a point on the characteristic represents the rate of consumption of power by the device at that point. In the case of a passive element this product is positive, and in the case of a battery or generator supplying power it is negative. In a number of the characteristics that have been encountered (tetrode, Fig. 7.13; arc, Figs. 8.5 and 8.9) the characteristic lies in the positive quadrant, but its slope over a certain limited region is negative. Small oscillations about a point in

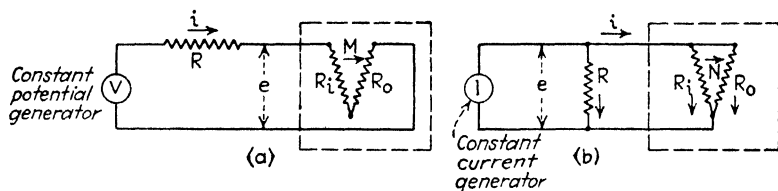



FIG. 15.10.—Schematic representation of the two types of unstable circuits. Type a, series- or current-controlled circuit; type b, shunt- or potential-controlled circuit.

this region represent a negative consumption or power output by the device as in the case of oscillations along the load line of a triode diagram. This, of course, requires the presence of an internal or external source of power and really represents a conversion of direct- into alternating-current power. As pointed out in Sec. 5.6, a negative slope of the over-all characteristic represents instability. Thus it is necessary to know the characteristic of the external circuit as well as that of the device itself before the action of the circuit can be predicted. In the following discussion the existence of a prime power source will be assumed, though it may not be specifically mentioned in every instance.

Since the power available in any circuit is finite, the characteristic can have **a** negative slope over only a limited region. But this region may be bounded in two ways, *i.e.*, by points of either zero or infinite slope as indicated in Fig. 15.12. Both types are met in practice (the arc is of type **a**, and secondary emission in the tetrode induces type **b**) and though the curves can be rendered of the same form by an interchange of the axes, it is more instructive physically to consider the two types separately.[1] Devices having these types of characteristics can be represented schematically by the three-terminal networks in the dashed

[1] HEROLD, *Proc. I.R.E.*, **23**, 1201 (1935).

boxes of Fig. 15.10. The characteristic $i = f(e)$ is presented by the two terminals that emerge. For type **a** the net is in series and M, which is characteristic of the device, is the emf. induced across R_0 per unit current in R_i. It differs from a coefficient of mutual inductance in that the reactions of R_i and R_0 are not equal and opposite (hence the necessity of a power supply which is not explicitly indicated). The emf. thus induced could be in either sense, but instability can be brought about

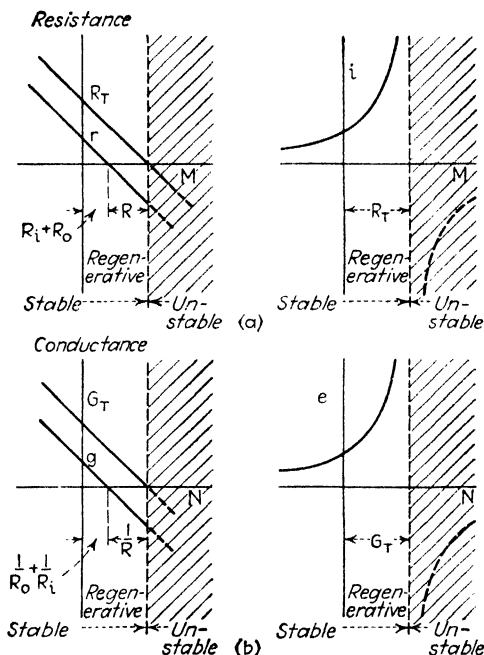



FIG. 15.11.—Type **a**, current-controlled circuit. Type **b**, potential-controlled circuit.

only if it is in such a sense as to compensate for the iR losses. The equation for the circuit is

$$V = (R + R_i + R_0 - M)i = (R + r)i = R_t i \qquad (15.13)$$

where $r = e/i = (R_i + R_0 - M)$ is the effective dynamic input resistance of the two-terminal device. M is considered as the independent variable and r, R_t, and i are plotted in the upper portion of Fig. 15.11. When r is less than zero, the losses in the device itself, $i(R_i + R_0)$, are more than compensated by the power it supplies, but owing to the external circuit the total losses are positive. However, when the total resistance R_t is less than zero, all the losses are more than supplied by the source of power associated with the device and the circuit is unstable. At the point $R_t = 0$ a finite current can flow even if the external potential V is reduced to zero. Thus the circuit becomes self-supporting when M exceeds

$$R + R_i + R_0$$

This point is predicted by the simple analysis, but the theory cannot be carried further without assuming a special form for the characteristic of the device.

Since the unidirectional interaction M determines the induced emf. per unit current, this type of device is said to be current-controlled. M is dependent on the various physical factors that determine the nature of the device, but as defined it is a linear function of the dynamic resistance r that the device presents. The characteristic associated with the type **a** device is shown at the left in Fig. 15.12, and beneath it is plotted the dynamic resistance or slope of the characteristic as a function of

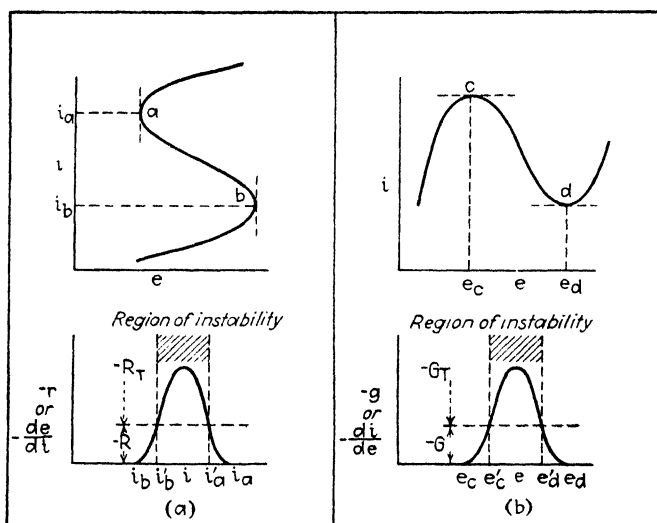



FIG. 15.12.—Typical characteristic curves in the neighborhood of instability. At left, type **a** characteristic, at right, type **b** characteristic.

the controlling current. From this point of view instability occurs in that region where r is negative and greater in magnitude than R. It is bounded by two values of the current i'_b and i'_a. The maximum amplitude of the alternating-current wave that can be produced is of the order of $(i'_a - i'_b)$. Considering the variation of the external resistance R, it is seen that as R is decreased, the inception of instability occurs when the horizontal line of ordinate $-R$ is tangent to the curve of r. At this point the amplitude of oscillation is infinitesimal, and as R is decreased, the amplitude increases approaching approximately $(i_a - i_b)$. Owing to the essential nonlinearity of the device harmonics are invariably associated with the output, though they may be greatly reduced by the use of resonant circuits and filters and by restricting the range of operation.

The characteristic of type **b** in Fig. 15.12 can be obtained from a

circuit represented schematically at the right in Fig. 15.10. The elements are in parallel and they may be considered to receive power from a constant-current generator. The unidirectional interaction N here takes the form of a current induced in R_0 per unit emf. across the circuit or across R_i. In terms of conductances (the conductance G is the reciprocal of the resistance R) the analysis of this circuit is formally the same as that of the preceding type **a**. Since the total flow of current to a junction is zero, the current supplied by I is equal to the flow through the parallel branches including specifically the induced current Ne. For instability to arise this latter must be in the assumed sense of I or opposite to that through the resistances. The equation for the circuit is thus

$$I = (G + G_i + G_0 - N)e = (G + g)e = G_t e \qquad (15.14)$$

where $g = i/e = (G_i + G_0 - N)$ is the effective dynamic input conductance of the two-terminal device. This is the reciprocal of the effective input resistance. Thus the discussion is exactly analogous to the preceding one with the interchange of e and i. This is evident since the characteristics are of exactly the same type if this interchange is performed. The lower portion of Fig. 15.11 indicates the regions of stability and instability as a function of N. Also, the right-hand portion of Fig. 15.12 represents this type of characteristic and the dependence of instability and amplitude of oscillation on the relation between g and the conductance of the external circuit, G. From this figure it is evident that instability and oscillations set in for small values of the external conductance or large values of the external resistance. This is the predominant practical distinction between the two types of circuit. For the type **a** circuit the external resistance must be small for instability; thus a series resonant circuit which presents a low resistance for the resonant frequency is of the type to induce instability and generate oscillations of approximately the resonant frequency. Whereas for the type **b** characteristic the external resistance must be large, *i.e.*, a parallel resonant circuit which presents a high resistance at the resonant frequency should be used for the load R to induce instability and generate oscillations of approximately the resonant frequency. The amplitude of the potential oscillations is of the order of $(e'_d - e'_c)$ of Fig. 15.12. Since the current Ne induced by the interaction N is determined by the potential e across the terminals, this type of device is known as potential-controlled.

The principal use of circuit elements having negative dynamic resistances is the production of an alternating current from a direct-current power source. However, these devices are also used to reduce the effective resistance of a circuit to a small positive value. In this case r is less in absolute magnitude than R and the effective circuit resistance is $R_e = R - r$. This service is known as *regeneration* (Sec. 15.5). A

mutual impedance between the plate and grid circuits transfers power from the former to the latter so that a portion of the grid circuit losses are supplied by the plate battery. Consider, for instance, that the cathode and grid are connected by a parallel-tuned circuit. An increase in the circulating power corresponds to an increase in the potential between cathode and grid. This increase in e_g results in a corresponding increase in i_p and an effective amplification constant exceeding that of the non-regenerative case. Since regeneration reduces the effective resistance it has the effect of increasing the effective Q of the resonant circuit. By this means the Q values of ordinary circuits can be made to approach those of quartz crystals, but they are less satisfactory for frequency discrimination as a further small decrease in effective resistance results in instability and oscillation.

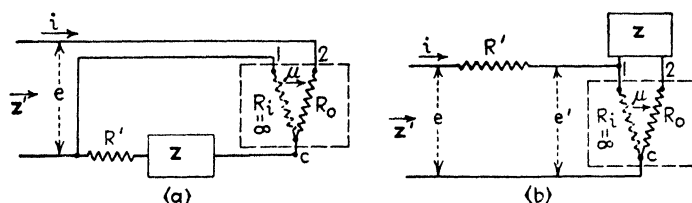



FIG. 15.13.—Circuits for obtaining a negative impedance or reactance. Type a, negative reactance; type b, negative reactance.

This type of circuit may also be used to produce reactances of unusual types. Both positive and negative (inductive and capacitative) reactances are familiar, but the ordinary inductive reactance varies as the first power of the frequency and the capacitative reactance as the inverse first power of the frequency. These relations may be reversed by means of circuits of the type of Fig. 15.13. The portion of the circuit having a negative dynamic resistance is in this case a vacuum tube with an effectively infinite input resistance R_i. Instead of M and N the circuits may be discussed by means of the amplification constant μ, which is the ratio $(\partial e_2/\partial e_1)_i$. This must be positive and greater than unity (opposite sign to the ordinary μ_p). From inspection the equation for the circuit of type **a** is

$$e = (R' + R_0 + \mathbf{z})i - \mu(R' + \mathbf{z})i$$

or

$$\mathbf{z}' = [R_0 - (\mu - 1)R'] - (\mu - 1)\mathbf{z}$$

Thus, if R' is chosen to have such a value that the first term vanishes, $\mathbf{z}'$ is an impedance of the same dependence on ω as $\mathbf{z}$ but of the opposite sign and greater in magnitude by $(\mu - 1)$. In the case of circuit **b**

$$e' - \mu e' = (R_0 + \mathbf{z})i$$

or

$$\frac{e}{i} = \mathbf{z}' = R' + \frac{e'}{i} = \frac{R_0 - (\mu - 1)R'}{1 - \mu} - \frac{\mathbf{z}}{\mu - 1}$$

Thus, if R' is so chosen that the first term vanishes, $\mathbf{z}'$ has the same dependence on ω as $\mathbf{z}$ but is of the opposite sign and smaller in magnitude by the factor $(\mu - 1)$. Without restriction $\mathbf{z}$ can be a pure reactance since the resistive component can be absorbed in R' or R_0. Thus a positive or inductive reactance, which, however, is inversely proportional to the frequency, can be produced if $\mathbf{z}$ is a capacity (a direct-current path must in general be provided) and a negative or capacitative reactance proportional to the frequency is produced if $\mathbf{z}$ is an inductance. The negative-reactance circuit can be combined with a positive reactance of like dependence on ω to form a circuit which has no frequency dependence, though there may be a phase shift dependent on ω. Also if a negative impedance $-\mathbf{z}_2$, equal in magnitude but opposite in sign to $\mathbf{z}_1$, is connected in the circuit of Fig. 15.14 the input impedance $\mathbf{z}'$ is given by $\frac{\mathbf{z}_1\mathbf{z}_2}{\mathbf{z}_4}$. Since all three of these impedances can depend on ω to the power 1, 0, or -1 an impedance can be constructed in this way which will depend on ω to any power from -3 to $+3$. Following this general principle more complex circuits can be devised for which the impedance will depend on any positive- or negative-integral power of ω.[1]

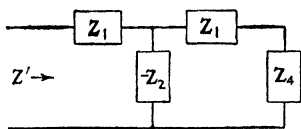



FIG. 15.14.—Basic circuit for the production of impedances dependent on positive- or negative-integral powers of the frequency.

15.7. Oscillators. *Internal Coupling or Negative Dynamic Resistance.*—An oscillator is in general any nonrotating device for the spontaneous conversion of direct-current into alternating-current power. The arc, glow discharge, and tetrode on the region of reverse plate current due to secondary emission are examples of the type in which internal processes are responsible for the falling characteristic presented by the terminals. The arc and glow discharge have characteristics of type **a** and their action is essentially current-controlled. However, the simple arc characteristic does not completely determine its electrical characteristics for these are also affected by the temperature and pressure in the various regions and the latter are in general variables as well. Arc phenomena are not sufficiently well understood to attempt an analysis of them here, but it may be mentioned that there is a lower limit and an upper limit to the frequencies of the sinusoidal oscillations that can be generated owing to the finite times associated with alterations in the arc processes. An ordinary carbon arc shunted by a series resonant circuit

[1] VERMAN, *Proc. I.R.E.*, **19**, 676 (1931).

can be made to oscillate strongly in the region of audio frequencies. The pulsations of the arc itself which are induced by the varying current make the arc act as an acoustic source and the frequency of oscillation may be determined by the note that is emitted. Secondary-emission devices (Sec. 7.3) generally present the type **b** characteristic and oscillations may be induced by the high resistance of a parallel resonant circuit. One type of circuit employs a triode, the grid being positive with respect to the cathode by a potential E_c and the plate also positive at a direct-current potential E_b, where E_b is less than E_c. The parallel-tuned circuit is in the lead from the plate to the battery. Secondary electrons emitted by the plate are drawn from it to the grid, resulting in a flow of current in the reverse sense in the plate circuit and the negative dynamic resistance associated with instability. The use of a tetrode with the first grid at a negative potential, the other connections being the same, is, however, preferable as the cathode current and grid dissipation are thereby limited to more reasonable values. This type of oscillatory circuit dependent on secondary emission is known as the *dynatron*. The efficiency of secondary emission is dependent on the particular nature of the plate surface and this may change rapidly under electron bombardment. Therefore these devices are in general somewhat unreliable.

A figure of merit of the type **a** circuit is the greatest negative dynamic resistance presented by the characteristic. For oscillations will be induced by a series resonant circuit if its resonant resistance (R) is less than this critical value. The negative dynamic resistance of the arc is in general rather low and a high-Q series circuit must be employed for the generation of oscillations. On the other hand, the figure of merit of the type **b** circuit is the greatest negative dynamic conductance presented by the characteristic. For this must exceed the effective dynamic conductance of the external circuit for the production of oscillations. Since the effective conductance of a parallel-tuned circuit at resonance is RC/L, where R is the resistance of the inductance, the condition for the generation of oscillations is $| -g | > RC/L$, where $| -g |$ is the magnitude of the negative dynamic conductance. If g is small, the Q of the oscillatory circuit must be large. For the dynatron g is of the order of 40 micromhos (r of the order of 25,000 ohms). Thus, if the inductance has a resistance of 10 ohms, the Q value of the resonant circuit must be greater than 50 for the production of oscillations.

Direct Coupling or Negative Dynamic Transconductance.—The ordinary vacuum tube is essentially a three-terminal device. The tube may contain a number of elements but except in certain cases of modulation and demodulation the potentials of all but two elements are constant, *i.e.*, an alternating-current potential appears between only two pairs of electrodes. If the transconductance between two elements is negative,

interaction between their respective circuits may result in an effective negative dynamic resistance and instability. The ordinary grid-plate transconductance of a triode is positive, but under certain conditions it may be made negative or the transconductance between another pair of elements may be used. One of the most satisfactory circuits[1] is that illustrated schematically in Fig. 15.15 in which the negative transconductance between the third and second grids of a pentode is employed. The first grid is at approximately the potential of the cathode (small variations in its potential may be used to control the magnitude of oscillation) and the plate which is maintained at a small constant positive potential merely collects the electrons which have passed through the grid systems. These electrodes play no role in the alternating-current portion of the circuit. The third grid is the negatively biassed control electrode

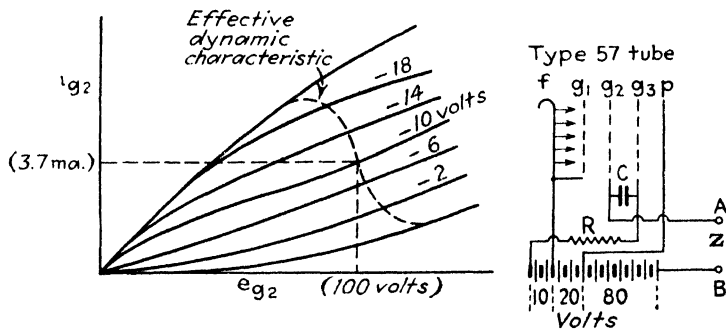



FIG. 15.15.—Negative transconductance oscillator.

and the second grid acts essentially as the plate. If g_3 becomes more negative, certain electrons which would previously have passed through g_2 and reached p are returned to the region of influence of g_2 and collected by that electrode. Thus a negative transconductance exists between these elements as illustrated graphically by the solid and dashed curves of the figure. The solid curves are the g_2 characteristics for various values of the potential of the control electrode g_3. If these electrodes are connected together (a condenser forming a low-resistance alternating-current path), the potential increments of these grids are equal, *i.e.*, $e_{g3} = e_{g2}$. Starting from the quiescent point this condition defines the dashed curve of the figure which is the dynamic characteristic. This dynamic transconductance is seen to be negative; hence a negative dynamic resistance is presented by the terminals AB. The condenser C merely plays the role of connecting the two electrodes together; its value is not critical, but if it is too large, "blocking" and relaxation oscillations will ensue. The resistance R, of the order of 10^6 ohms, isolates g_3 from the battery circuit. It should be noted that the amplifica-

[1] HEROLD, *loc. cit.*

tion constant between these electrodes as represented by this type of dynamic characteristic is essentially positive and hence this type of element can be utilized in the circuits of Fig. 15.13.

The ordinary triode may exhibit a negative transconductance under various circumstances. A small amount of residual gas frequently gives rise to this phenomenon and it may also be observed if the space current is limited by cathode emission. Furthermore, for oscillations of very short period, comparable to the time of flight of the electrons through the grid structure, the effective transconductance may change sign. In the dynatron region of the tetrode characteristic the transconductance is also negative. However, the most satisfactory and reliable circuit is that of Fig. 15.15 employing one of the standard low-power pentodes. The constants associated with the 57 in the region of the direct-current potentials specified in the figure are

$$r_{g2} = \frac{\partial e_{g2}}{\partial i_{g2}} = 40{,}000 \text{ ohms} \qquad \text{and} \qquad s_{32} = \frac{\partial i_{g2}}{\partial e_{g3}} = -310 \times 10^{-6} \text{ mho}$$

The product of these yields the amplification constant for these electrodes as 12.4. The variable current to grid 2 is given to the linear approximation by

$$i_{g2} = \frac{\partial i_{g2}}{\partial e_{g2}} e_{g2} + \frac{\partial i_{g2}}{\partial e_{g3}} e_{g3}$$

For direct coupling $e_{g2} = e_{g3}$; hence

$$\frac{i_{g2}}{e_{g2}} = (25 - 310) \times 10^{-6} = -285 \text{ micromhos}$$

This is seen to be considerably greater than for the dynatron previously quoted, and as a consequence it is not necessary to employ a resonant circuit with as high a Q for the production of oscillation.

Positive Transconductance and Reverse Phase Coupling.—The ordinary triode has a positive transconductance, but the external circuit coupling the grid and plate may be of such a nature that a phase reversal takes place which effectively converts a positive dynamic resistance into a negative one. This may be done in various ways the most common ones being capacitative coupling between grid and plate or inductive coupling of the proper sense between the two circuits. The great majority of oscillatory circuits are of the reverse-phase coupled type and a few representative ones will be discussed. Consider first the type that depends on capacitative coupling between plate and grid. The fundamental circuit involved is shown in the upper portion of Fig. 15.16. The tube capacities are shown explicitly and C_{pg} may be augmented by an external capacity C' which will, however, be included in C_{pg} in the analysis. Provided no alternating potential is applied to

the grid circuit in series with an impedance the plate and grid circuits are completely symmetrical for an alternating-current component and a parallel or shunt analysis is the most obvious one. If the tube capacities are neglected Eq. (7.6) applies.

$$\mathbf{i}_p = s_p\mathbf{e}_g + k_p\mathbf{e}_p \qquad (7.6)$$

Or writing $\mathbf{i}_p = \mathbf{y}_l\mathbf{e}_l$, where $\mathbf{y}_l$, which is equal to $1/\mathbf{z}_l$, is the load admittance and $\mathbf{e}_l = -\mathbf{e}_p$ is the alternating potential appearing across the load

$$s_p\mathbf{e}_g = k_p\mathbf{e}_l + \mathbf{y}_l\mathbf{e}_l \qquad (7.11)$$

The left side represents a current delivered by a resistanceless generator in the plate circuit and the right side represents the sum of the separate

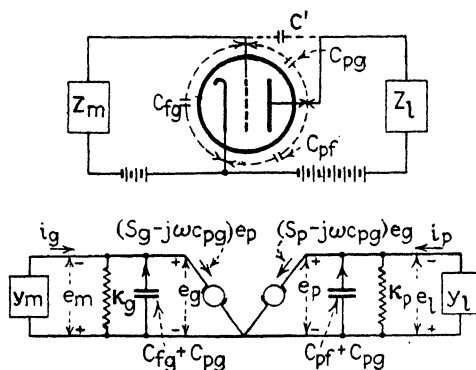



FIG. 15.16.—Schematic analysis of an oscillatory vacuum-tube circuit. (a) Schematic vacuum-tube circuit. (b) Equivalent parallel circuit.

components flowing through the plate resistance and load impedance. The effects of the interelectrode capacities can be readily included in this scheme. In the first place C_{pf} forms an additional shunt carrying a current $j\omega C_{pf}\mathbf{e}_l$. The capacitative current from the grid to the plate is $(\mathbf{e}_g - \mathbf{e}_p)j\omega C_{pg}$ or $(\mathbf{e}_g + \mathbf{e}_l)j\omega C_{pg}$. When these currents are added to the right side of Eq. (7.11) and the terms containing $\mathbf{e}_g$ collected on the left

$$(s_p - j\omega C_{pg})\mathbf{e}_g = [k_p + \mathbf{y}_l + j\omega(C_{pg} + C_{pf})]\mathbf{e}_l$$

This equation is seen to correspond to the right-hand portion of the equivalent circuit in Fig. 15.16. $\mathbf{e}_l$ could, of course, be expressed in terms of $\mathbf{y}_l$ and $\mathbf{i}_p$ if desired. An exactly analogous equation can be developed for the grid circuit starting from Eq. (7.6′)

$$(s_g - j\omega C_{pg})\mathbf{e}_p = [k_g + \mathbf{y}_m + j\omega(C_{pg} + C_{fg})]\mathbf{e}_m$$

The assumption of a linear circuit with constant values of the s's and k's is admittedly inadequate to describe the oscillatory state. But

if these two simultaneous equations permit of a solution in terms of finite values of $\mathbf{e}_g$ and $\mathbf{e}_l(\mathbf{e}_g = -\mathbf{e}_m$ and $\mathbf{e}_l = -\mathbf{e}_p)$, the network losses must be completely made up by the direct-current power sources and oscillations of a finite amplitude can be maintained. Eliminating the potentials yields the condition

$$(s_p - j\omega C_{pg})(s_g - j\omega C_{pg}) = [k_p + \mathbf{y}_l + j\omega(C_{pg} + C_{pf})][k_g + \mathbf{y}_m + j\omega(C_{pg} + C_{fg})]$$

The restrictions implied on the real and complex parameters by this equation determine the condition of instability and the frequency of the oscillations generated. Neglecting grid conduction current, which implies $s_g = 0$ and $k_g = 0$, and solving explicitly for $\mathbf{y}_m = g_m + ib_m$

$$-\mathbf{y}_m = -g_m - ib_m = \frac{j\omega C_{pg}(s_p - j\omega C_{pg})}{k_p + \mathbf{y}_l + j\omega(C_{pg} + C_{pf})} + j\omega(C_{pg} + C_{fg})$$
$$= \frac{\mu C_{pg} + (C_{fg} + C_{pg})(r_p\mathbf{y}_l + 1) + j\omega r_p(C_{fg}C_{pg} + C_{fg}C_{pf} + C_{pg}C_{pf})}{r_p(C_{pg} + C_{pf}) - \frac{j}{\omega}(r_p\mathbf{y}_l + 1)}$$

If the real part of the expression on the right is negative and exceeds g_m, the circuit will be unstable. On rationalization it can be seen that the real part can be negative only if the imaginary part of $\mathbf{y}_l$ is negative. This implies an inductive reactance for the load and the input conductance, *i.e.*, the real portion of the expression on the right is found to be negative in this case if

$$\mu L_l > r_pR_lC_{pg} + (R_l^2 + \omega^2L_l^2)[\mu(C_{pg} + C_{pf}) + C_{pg}] \qquad (15.15)$$

Here R_l and L_l are the resistance and inductance of the load, respectively. Thus, if the inductance of the load lies between the two values defined by this inequality, the circuit may be unstable and oscillations may arise. This situation frequently occurs when it is not desired and may be remedied by an increase in R_l or a decrease generally in L_l A determination of the actual frequency of oscillation requires a complete solution of the equation of condition.

A vacuum tube with a parallel-tuned circuit in the grid and another in the plate circuit will generate oscillations when properly adjusted owing to the capacitative interaction between the plate and grid. The frequency of oscillation is primarily determined by the product of the inductance and capacity in the grid circuit. Since it has been seen that the plate load must present an inductive reactance, this circuit must be tuned to a somewhat higher frequency than the natural frequency of the grid circuit. Such an oscillator is known as the *tuned-plate tuned-grid* type. A number of interesting principles are illustrated by the oscillator circuit of Fig. 15.17. Here the second grid plays the role of

the plate in the ordinary oscillator; the third grid acts as a screen to reduce any reaction of the plate circuit on the oscillator proper. The space current to the plate is essentially modulated by the oscillating potentials of the first two grids and the plate output is then amplified and used. The arrangement is known as *electron coupling* and is frequently used to isolate the oscillatory portion from the succeeding meshes of a circuit. A quartz crystal, which is essentially a series resonant element, is shown in the grid circuit.[1] The oscillations produced are very nearly those corresponding to the natural frequency of the crystal. Owing to the very high Q-value of this element, a very small change in frequency will result in a large change in the effective parameters which it presents to the circuit. Since the conditions of oscillation may be made to depend critically on these parameters, the frequency of oscillation may be held within very narrow limits by this type of circuit. It is the one which is generally employed for the production of constant frequencies. If circuit elements of high quality are used and the circuit is carefully adjusted, it is superior to the best clocks for the measurement of time intervals. The auxiliary condenser C' is for adjusting the frequency to any predetermined value in the immediate neighborhood of the natural frequency of the crystal. As in the case of most oscillator circuits, the mean grid potential is determined by the $\mathbf{I}_c r_c$ drop in a resistance in series with the grid. This method of obtaining grid bias makes the oscillator self-starting and the bias self-adjusting. The alternating current that the ordinary crystal a few square centimeters in area can carry safely is of the order of 0.05 amp., but for the production of constant frequencies it should be kept to a much smaller value. The alternating current is of the order of the direct-current component of the grid current; hence this consideration and the mean operating grid potential determine the order of magnitude of r_c. C_b is merely a condenser for by-passing the alternating current across the battery. From the previous discussion it is evident that the reactance presented by the tuned LC circuit must be positive; hence it must be tuned to a somewhat higher frequency than that of the crystal. The circuit has been analyzed in detail by Wheeler[2] who has shown that if certain conditions are satisfied, the frequency of oscillation depends primarily on the crystal itself and is relatively independent of variations in other circuit parameters. These conditions are:

FIG. 15.17.—Electron-coupled crystal oscillator.

[1] For a description of the magnetostriction oscillator for performing the same function at lower frequencies see Pierce, *Proc. I.R.E.*, **17**, 42 (1929).

[2] WHEELER, *Proc. I.R.E.*, **19**, 627 (1931).

(*a*) the Q of the LC circuit should be as small as consistent with stable oscillations; (*b*) the plate resistance should be as high as consistent with stable oscillation; (*c*) the capacity C_{gf} (including that of the holder and C') should be close to, but not in excess of, $(\mu - 1)C_{gp}$; (*d*) the capacity C should be adjusted for maximum plate current.[1]

The other method of reverse phase coupling is that employing an inductive reaction between the plate and grid circuits. This is the type most widely used for power oscillators. The resonant circuit may be in either the plate or grid circuit, as shown in Fig. 15.18, or common to the two, as in Fig. 15.19. The latter type is known as the *Hartley oscillator.* The general nature of the linear-circuit analysis is the same for all and if the grid current is neglected (which is legitimate to a first approxi-

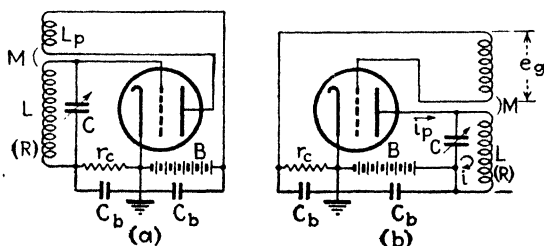



FIG. 15.18.—(*a*) Tuned-grid oscillator. (*b*) Tuned-plate oscillator.

mation), only two meshes are involved and the solution is relatively simple. Consider for example the *tuned-plate oscillator* of Fig. 15.18.

$$\mu \mathbf{e}_g = \left(r_p - \frac{j}{\omega C}\right)\mathbf{i}_p + \frac{j}{\omega C}\mathbf{i} \qquad \text{and} \qquad \mathbf{e}_g = Mj\omega\mathbf{i}$$

or

$$0 = \left(r_p - \frac{j}{\omega C}\right)\mathbf{i}_p - j\left(\omega M\mu - \frac{1}{\omega C}\right)\mathbf{i}$$

and secondly

$$0 = \frac{j}{\omega C}\mathbf{i}_p + \left[R + j\left(\omega L - \frac{1}{\omega C}\right)\right]\mathbf{i}$$

For these equations to have a solution other than $\mathbf{i}_p = \mathbf{i} = 0$ the determinant of the coefficients must vanish. Equating the real and imaginary portions to zero yields

$$M = \frac{(Rr_pC + L)}{\mu} \qquad \text{and} \qquad \omega = \omega_0\left(1 + \frac{R}{r_p}\right)^{1/2}$$

[1] The use of a high-Q mechanical element is the most satisfactory method of generating constant frequencies. The variation in the effective tube parameters is the factor which is largely responsible for frequency fluctuation in the ordinary type of circuit. This source of frequency instability can be greatly reduced by special circuit designs that have been derived by Llewellyn [*Proc. I.R.E.*, **19**, 2063 (1931)].

where $\omega_0 = 1/(LC)^{1/2}$. The condition involving M represents the verge of instability, and oscillations will occur if M is of this value or greater. Actually, of course, r_p is only approximately a constant and the equation for M may be taken as determining the effective value of r_p in the oscillatory state. It is evident that M must be in the proper sense for regeneration if oscillations are to be produced. This is positive in accordance with the convention used above. The equation for ω shows that the frequency of oscillation is greater by the factor in the bracket than the natural frequency of the resonant circuit. A rough criterion for deter-

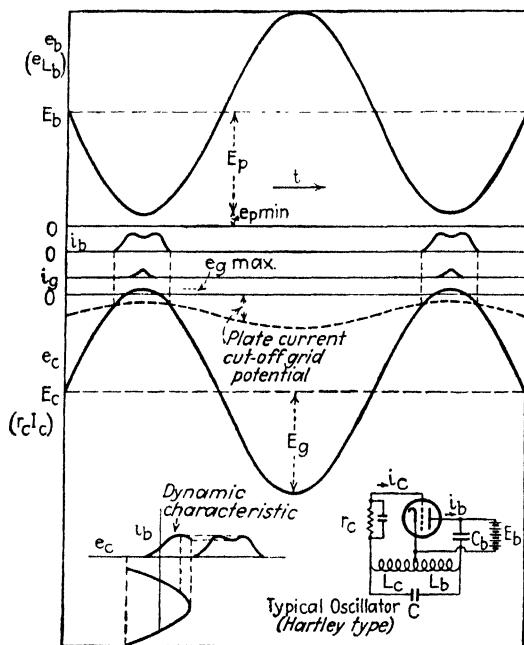



FIG. 15.19.—Phase relations between the plate and grid circuits of an oscillator.

mining the magnitude of M is that for oscillations to persist it is generally necessary that the alternating emf. induced in the grid circuit be of the order of three times the cutoff grid potential corresponding to the potential of the plate battery or generator.

The instantaneous potentials of the elements and the currents to them are indicated as functions of the time in Fig. 15.19. The large circulating current in the tank circuit makes the potentials e_g and e_p approximate closely to sine waves. The total plate potential e_b is the battery potential E_b plus this alternating component. The total grid potential is the mean grid potential E_c (determined by the mean grid current and the grid resistance r_c) plus an oscillatory component proportional to e_p, but in the opposite phase, *i.e.*, proportional to $-e_p$.

Plate current flows during the portion of the cycle in which the grid potential exceeds the effective cutoff value corresponding to the instantaneous plate potential. The plate current during the interval is in general not a portion of a sine loop owing to the curvature of the dynamic characteristic. The central minimum shown in the figure is largely due to the fact that the current to the electrodes is limited by cathode emission and part of the current flows to the grid when this element is positive. To avoid excessive grid losses the maximum positive excursion of the grid ($e_{c\ \max}$) should not exceed about 80 per cent of the minimum instantaneous plate potential ($e_{b\ \min}$). This type of oscillator is essentially a self-excited class C amplifier and the discussion of that device is largely applicable to the oscillator. The total losses (exclusive of cathode heating) must, however, be supplied by the direct-current plate source. These losses include the loss at the grid and plate as well as that lost in the grid resistor r_c and that supplied to the load. The power equation may be written

$$E_bI_b = \frac{1}{\tau}\int e_ci_c\,dt + \frac{1}{\tau}\int e_bi_b + r_cI_c^2 + \frac{I_pE_p}{2}$$

Power supplied	loss at grid	loss at plate	grid resistor loss	a.-c. power supplied to load

Since the plate circuit efficiency is always greater than the efficiency of the plate and grid circuits together, more power can be obtained from a class C amplifier than from an oscillator for the same external circuit conditions.

It is evident that the plate loss can be reduced by reducing the time of conduction, *i.e.*, by increasing r_c. The output can be maintained by increasing E_b. Thus the efficiency can be increased up to a certain point. If r_c is made too large the behavior of the circuit may be completely altered. The time constant of r_c and the condenser may become so large that the mean negative direct-current potential of the grid is too great for oscillation. Oscillation ceases until a sufficient number of electrons have leaked off through r_c to permit space current to flow. Inception of grid current again induces cutoff and nonsinusoidal oscillations thus occur. These are known as *relaxation oscillations* and their period is determined by the time constant of the grid circuit. On the other hand, an increase in r_c and the high positive grid potential peaks that are then necessary for very brief portions of the cycle may give rise to an entirely different phenomenon. The electrons may strike the grid with sufficient energy to produce appreciable secondary emission. If the plate is more positive than the grid, these electrons are drawn to the plate and a positive current effectively flows to the grid. As the grid becomes more positive the space current rises rapidly. This results in

excessive heating of the electrodes, liberation of gas, destruction of the cathode surface and frequently of the tube itself.

The design and adjustment of an oscillator are largely empirical and based on previous practice aided by graphical analysis. In small oscillators the grid resistance r_c is varied throughout the range from 10^3 to 10^5 ohms until satisfactory operation is achieved. In power oscillators it is adjusted till the peak positive grid potential is about 80 per cent of the minimum instantaneous plate potential. The design of the tank or load circuit is also important for satisfactory operation. As in the class C oscillator, it plays the role of a flywheel which is kept in oscillation by the periodic impulses received from the plate battery. If the total effective Q of this circuit (including reflected load) is large, the circulating current is large in comparison with the current impulses received and the flywheel action is very effective. However, the power delivered to the load during the brief period of flow of plate current is small. On the other hand, if the Q of the circuit is small the circulating current is small and the apparent amplitude decreases during a cycle, leading to poor wave form. The result of experience is that the optimum value of the Q of this circuit (inclusive of reflected load) is of the order of 4π or somewhat greater. If I is the effective value of the circulating current in the parallel-resonant tank circuit and E the effective value of the potential appearing across it

$$E \cong \omega LI \qquad \left(E \text{ is approximately } \frac{E_B}{\sqrt{2}}\right)$$

Furthermore, the power delivered is $P = I^2R$, where R is the resistance of the inductance, including the reflected load coupled to the oscillatory circuit. Hence

$$Q = \frac{\omega L}{R} \cong \frac{EI}{P}$$

Thus Q is also the ratio of the mean circulating power to the power consumed by the load. The order of magnitude of suitable inductance and capacity can be determined from the relations

$$E \cong \omega LI \cong \frac{I}{\omega C} \qquad \text{and} \qquad Q = \frac{EI}{P} = 4\pi$$

yielding

$$L = \frac{E^2}{8\pi^2 \nu P} \qquad C = \frac{2P}{E^2 \nu} \qquad (2\pi\nu = \omega)$$

These values are to be considered as only approximate. For a constant E and constant frequency the circulating current in the tank circuit itself, which is $E_b\sqrt{C/L}$, is inversely proportional to $L^{1/2}$ and directly proportional to $C^{1/2}$. Thus the circulating current is increased by decreasing L and increasing C from the values given by the above equations if E

and ν are held constant. This improves the wave form by increasing the ratio of circulating current to incremental current received from the tube, but also increases strictly tank-circuit losses.

Very High-frequency Oscillators.—An upper limit to the frequency of oscillation for the standard circuits that have so far been discussed is set by the physical characteristics of the circuit elements. The changes in the behavior of inductances, capacitances, and resistances were discussed in Sec. 13.3. These can no longer be considered as simple lumped constants leading to resonant combinations with a single characteristic frequency, but instead they resemble the transmission lines of Sec. 14.5 in having a whole spectrum of characteristic frequencies. Vacuum-tube structures also present reactive impedances that are functions of the frequency. Furthermore when the time taken for an electron to travel from cathode to plate becomes comparable with the period of oscillation, the preceding analyses of the tube's action no longer apply. The frequency range in which tube structures of the type so far considered are useful can be extended upward somewhat by decreasing electrode separations and increasing the applied potentials, but a practical limit is soon reached owing to construction difficulties and field emission. The use of frequencies in the range from 10^9 to 10^{11} cycles per second necessitates entirely different principles of tube design. Great advances were made in this field during the Second World War in connection with radar development. The electron motion is controlled in such a way by applied electric and magnetic fields as to excite the dominant modes of oscillation of certain metallic cavities associated with the tube structure (Sec. 16.5). These electromagnetic oscillations are then propagated down transmission lines or wave guides and ultimately radiated into space (Chap. XVI). The frequencies that can be generated by these techniques are comparable to certain characteristic atomic and molecular frequencies and in consequence provide a valuable tool in physical research.

There are two types of generators of these very high-frequency oscillations. The first is the *resonant cavity magnetron*, and the second is the *klystron*. The general principle of the magnetron was described in Sec. 9.2, and a section through the structure of a resonant cavity magnetron is indicated in Fig. 15.20. The anode consists of a copper block through which is bored a cylindrical cavity surrounding the axial cathode. The latter is supported from its ends and has an oxide surface for copious electron emission. The constant accelerating field for the electrons is radial, and the magnetic field is normal to the plane of the figure. These fields are so related that the electron trajectories are approximately tangential to the inner surface of the anode. Symmetrically disposed about the center are an even number of cylindrical cavities through the

block connecting with the central cavity by slots. The mode of oscillation indicated is that in which currents circulate in planes normal to the axis and in opposite senses in neighboring cavities, inducing opposite charges on the walls of the slots and segments of the anode surface. Alternate segments are strapped together at the ends of the anode block to promote this mode of oscillation. The electron trajectories are not simply calculable, but the magnitudes of the fields are such that the rotation of the electronic space charge is in synchronism with the oscil-

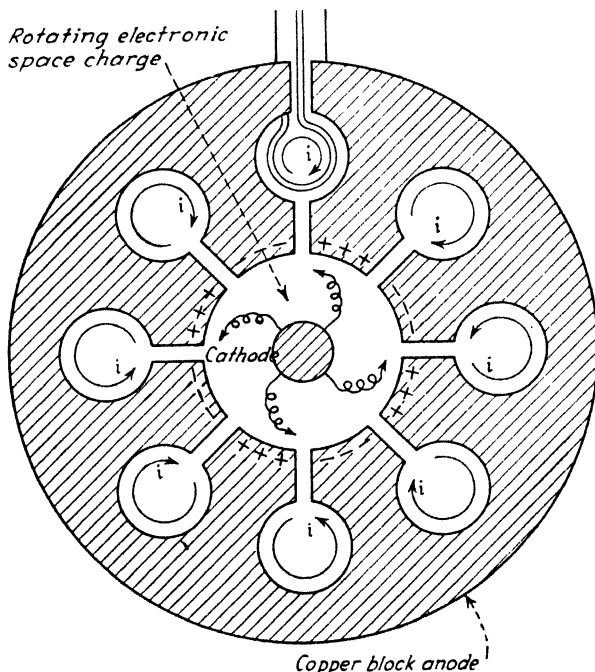



FIG. 15.20.—Schematic section through a resonant cavity magnetron.

latory mode excited and the energy of oscillation is supplied by the electron motion against the local tangential electric fields at the inner anode surface. A coupling loop and beginning of a transmission line is indicated at the top of the figure. The circulating current in the cavity induces a circulating current in the central conductor of the line, and the intercoupling of the eight cavities is so close that the energy is drawn from the entire oscillating system. The frequency is determined primarily by the dimensions of the cavity-slot structure, and the fields are adjusted to be in resonance with this. Magnetrons are not generally operated continuously, but peak powers of some hundreds of kilowatts can be generated by them at frequencies up to about 3×10^{10} cycles per second.[1]

[1] COLLINS, "Microwave Magnetrons," McGraw-Hill Book Company, Inc., New York, 1948.

The klystron is a tube in which oscillations are generated by means of the interaction of an electron beam and a cavity without the necessity of a magnetic field. Figure 15.21 represents schematically a section through the axis of such a tube defined by the electron beam. Except for the output line at O it may be thought of in space as a figure of revolution about the axis of this beam. The beam traverses the cavity C in the form of a disk having grids near its center. On the opposite side of the cavity it encounters a retarding field, which reflects it back through the cavity again. The circulating current indicated by i is excited by the

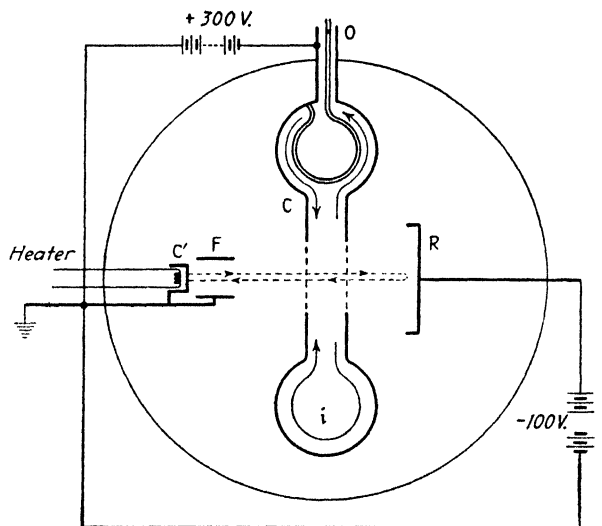



FIG. 15.21.—Schematic reflex klystron.

beam, which is in turn modulated by it in such a sense as to contribute to the oscillation. The resonant condition is achieved by varying the potential of the reflecting electrode until the time spent by the electrons in the retarding field region is such that they return through the cavity in the proper phase to contribute to the oscillation. In the following analysis all fields are assumed to be along the z axis, which is that of the beam. The potential difference V' between the grids caused by the oscillating current in the cavity is assumed to be small in comparison with the static accelerating potential V, and it is also assumed that the change in V' during an electron's transit between the grids is negligible. Then the energy of an electron on leaving C is $\frac{1}{2}mv_1^2 = e(V + V' \sin \delta)$, where δ is the phase of oscillation at transit. The electron then experiences a constant deceleration from R until it is brought to rest; then it reverses and returns through C. The time spent between C and R is easily calcu-

lated to be $t = 2mv/eE_r$, where e and m are, respectively, the electron charge and mass; v is its velocity on leaving C; and E_r is the retarding field. There is no interchange of energy on the average between the electron beam and the cavity the first time through, because the average value of $\sin \delta$ over a cycle vanishes. However, the modulation imposed on the electron beam during its first passage enables it to do a net amount of work against the cavity field the second time through. The phase at which an electron, which initially passed through at δ, traverses the cavity the second time is $\delta + \omega t$. This transit is in the reverse sense, and in consequence its energy on emerging from C is

$$\tfrac{1}{2}mv_2^2 = e[V + V' \sin \delta - V' \sin (\delta + \omega t)]$$

As it passed this point initially with the energy eV and the average of the second term vanishes, the average fraction of the beam energy gained by the cavity to sustain oscillation is

$$F = \frac{f}{2\pi}\int_0^{2\pi} \sin (\delta + \omega t)\, d\delta$$

where $f = V'/V$, and from the preceding discussion

$$t = \left(\frac{8mV}{eE_r^2}\right)^{\frac{1}{2}}\left(1 + \frac{f}{2}\sin \delta\right)$$

This is a well-known integral, which can be expressed in terms of the first-order Bessel function as

$$F = f \sin \alpha J_1\left(\frac{\alpha f}{2}\right)$$

where $\alpha = (8mv\omega R/eE_r^2)^{\frac{1}{2}}$. The quantity f has been assumed small in this argument, the sine term cannot exceed unity, and the maximum value of J_1, which occurs at $\alpha f = 3.68$, is 0.58. In consequence F is small and the device is relatively inefficient. However, it can be maximized by suitable choice of V and E_r, and it is found in practice that efficiencies of the order of 1 per cent are obtained. The frequency of oscillation is determined primarily by the cavity shape and dimensions, although E_r has a small effect upon it. Typical klystron power levels are from 2 to 5×10^{-2} watt, and they are used to produce continuous oscillations at low power levels.[1]

15.8. Simplified Nonlinear Oscillator Theory.[2]—A general analytical solution of even the simplest oscillator circuit would be extremely complex, but if certain radical

[1] HAMILTON, KNIPP, and KUPER, "Klystrons and Microwave Triodes," McGraw-Hill Book Company, Inc., New York, 1948.

[2] VAN DER POL, *Proc. I.R.E.*, **22,** 1051 (1934); BRAINERD and WEYGANDT, *Proc. I.R.E.*, **24,** 914 (1936); MINORSKY, "Introduction to Non-linear Mechanics," Edwards Bros., Inc., Ann Arbor, Mich., 1947.

simplifying assumptions are made, an approximate analytical solution can be obtained. Though the solution so obtained is not of great practical value, it yields interesting qualitative information in regard to the operation of certain oscillatory systems. It will be assumed that the characteristic of the unstable circuit element is a uniquely defined single-valued function of the independent variable. It will further be assumed

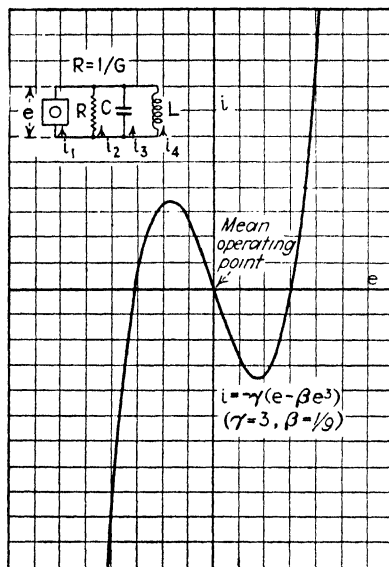



FIG. 15.22.—Type of characteristic assumed for oscillator analysis.

that the characteristic is a symmetrical cubic function and that the mean alternating-current operating point is the point of inflexion as shown in Fig. 15.22. Thus

$$e = -\rho(i - \alpha i^3) \qquad \text{(type } \mathbf{a} \text{ characteristic)}$$
$$i = -\gamma(e - \beta e^3) \qquad \text{(type } \mathbf{b} \text{ characteristic)}$$

The circuit analysis is of the same form for either type; a series circuit being employed for type **a** and a shunt circuit for type **b**. Taking the latter as an example, the sum of the currents in the diagrammatic circuit of Fig. 15.22 is equal to 0. Since

$$e = Ri_2 = L\frac{\partial i_4}{\partial t} = \frac{1}{C}\int i_3\,dt$$

and

$$i_1 + i_2 + i_3 + i_4 = 0$$
$$-\gamma(e - \beta e^3) + Ge + \frac{1}{L}\int e\,dt + C\frac{\partial e}{\partial t} = 0$$

Since t is the only independent variable, total derivatives may be used. Differentiating again to avoid the integral term the equation becomes

$$C\frac{d^2e}{dt^2} + [G - \gamma(1 - 3\beta e^2)]\frac{de}{dt} + \frac{1}{L}e = 0 \qquad (15.16)$$

The equation for the type **a** characteristic and series circuit is the same with i, L, R, ρ, α, and C substituted for e, C, G, γ, β, and L, respectively. This equation can be

simplified by the substitutions

$$\epsilon = \sqrt{\frac{L}{C}}(\gamma - G), \qquad x = \sqrt{\frac{3\beta\gamma}{\gamma - G}}e = \delta e, \qquad t' = \frac{t}{(LC)^{1/2}} = \omega_0 t$$

yielding

$$\frac{d^2x}{dt'^2} - \epsilon(1 - x^2)\frac{dx}{dt'} + x = 0 \tag{15.17}$$

The nature of the solution of this equation depends primarily on the sign of ϵ. If ϵ is negative, there is an oscillatory solution with positive damping. The coefficient of e in the substitution equation, $\delta = \sqrt{3\beta\gamma/(\gamma - G)}$, is a measure of the departure from the simple harmonic type. Thus the magnitude of ϵ may be taken as a measure of the damping and δ as a measure of the complexity of the wave form. If $\epsilon = 0$. the oscillations are sinusoidal of arbitrary amplitude and with a period $2\pi\sqrt{LC}$.

The more interesting cases are those for which ϵ is positive, for then the damping is negative and the amplitude of oscillation increases. Take the simplest case in which ϵ and δ are small. Owing to the latter assumption, the oscillations will be approximately sinusoidal, hence a solution of the form

$$x = u \sin t' \tag{15.18}$$

in which u is a slowly varying function of t', will be assumed. The second derivative of u with respect to t' and the product of the first derivative and ϵ will be considered small enough to neglect. Then

$$\begin{aligned}
\epsilon\frac{dx}{dt'} &= \epsilon u \cos t' \\
\frac{d^2x}{dt'^2} &= 2\frac{du}{dt'}\cos t' - u \sin t' \\
\epsilon x^2 \frac{dx}{dt'} &= \frac{\epsilon}{3}\frac{d(x^3)}{dt'} = \epsilon\frac{u^3}{3}\frac{d(\sin^3 t')}{dt'} \\
&= \epsilon\frac{u^3}{3}\frac{d[\frac{1}{4}(3\sin t' - \sin 3t')]}{dt} \\
&= \epsilon\frac{u^3}{4}\cos t' \qquad \text{(neglecting the harmonic term)}
\end{aligned}$$

Substituting these in Eq. (15.17) yields

$$2\frac{du}{dt'} - \epsilon u\left(1 - \frac{u^2}{4}\right) = 0$$

Multiplying through by u and employing the substitution $u^2 = 1/v$ yields

$$\frac{dv}{dt'} + \epsilon\left(v - \frac{1}{4}\right) = 0$$

This may be integrated immediately by separation of the variables to give:

$$v = \tfrac{1}{4}[1 + e^{-\epsilon(t' - t_0')}] \tag{15.19}$$

where t_0' is a constant of integration. From Eq. (15.19)

$$u = 2[1 + e^{-\epsilon(t' - t_0')}]^{-1/2}$$

and

$$e = 2\sqrt{\frac{\gamma - G}{3\beta\gamma[1 + e^{-\epsilon\omega_0(t - t_0)}]}}\sin \omega_0 t \tag{15.20}$$

This corresponds to a sine wave with an amplitude that increases with the time (owing to the t in the exponent of the denominator) until it reaches a maximum value of $\left[\frac{4(\gamma - G)}{3\beta\gamma}\right]^{1/2}$ after a very long time.

If the oscillations are not approximately sinusoidal, a solution of the differential equation can be obtained only by graphical or mechanical means. However, an interesting approximate solution for a case of this type is that of the two-stage resistance-capacity-coupled amplifier with the output returned to the input. A device of this type is known as a *multivibrator* and exhibits the relaxation type of oscillations. Consider the circuit of Fig. 15.23 and assume that the grid currents are negligible and that R_g is very much greater than R_p so that the current i is much smaller than the current i_p. Then, since the by-pass condenser C_b will be assumed very large, the

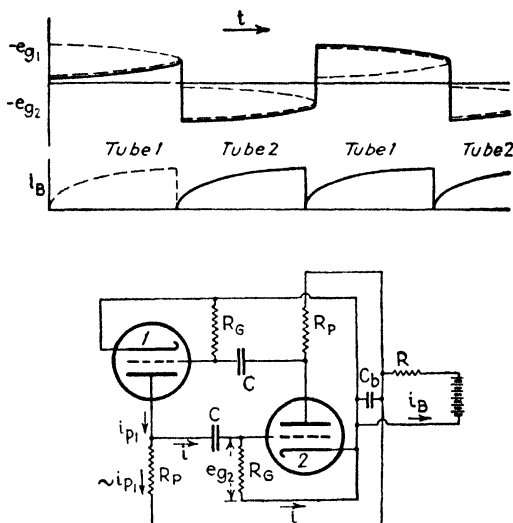



FIG. 15.23.—Typical relaxation oscillator or multivibrator.

sum of the potentials across R_p, C, and R_g is approximately zero. Assuming a cubic equation for the dependence of i_p on e_g

$$i_{p1} = s_p(1 + \beta e_{g1}^2)e_{g1}$$

and

$$R_p i_{p1} = \int \frac{i}{C} dt + R_g i$$

Writing $-e_{g2}$ for $R_g i$ and taking the derivative of the equation with respect to t

$$s_p R_p(1 + 3\beta e_{g1}^2)\frac{de_{g1}}{dt} + \frac{e_{g2}}{R_g C} + \frac{de_{g2}}{dt} = 0$$

There is, of course, an analogous equation for the other circuit. Assuming further that the circuit is symmetrical and that the variation in potential of one grid is equal and opposite to that of the other ($e_{g1} = -e_{g2}$), the two equations become identical and can be written

$$R_g C[(R_p s_p - 1) + 3R_p s_p \beta e_g^2]\frac{de_g}{dt} - e_g = 0$$

This may be integrated immediately to yield

$$a \log e_g^2 - b e_g^2 = \frac{2}{RC}(t - t_0) \tag{15.21}$$

where a is written for $(R_p s_p - 1)$ and b for $-3R_p s_p \beta$. t_0 is a constant of integration. If b is considerably smaller than a, this equation represents the type of curve shown by one of the dashed units of Fig. 15.23. This curve describes the variation in grid potential between points of discontinuity in the operation of the circuit. The general nature of the variation of the grid potentials and plate currents with time is shown by the solid curves in the upper portion of Fig. 15.23. The approach to the cyclic state is shown along the lower branch of a curve of the type of Eq. (15.21), but in the cyclic state it is only the portions most distant from the axis that are traversed.

The action of the circuit may be thought of as follows: The grids are biassed negatively by electrons reaching them from the cathode. The charge acquired by C leaks off through R_g and assuming this takes place somewhat more rapidly in tube 1 (owing to slight asymmetry), this tube becomes conducting first. The potential of the plate of this tube drops steadily as conduction increases. The grid of tube 1 traverses the upper branch of one of the curves and the plate current in that tube rises. The fall in plate potential reacting through C on the grid of tube 2 keeps it nonconducting during this interval. However, a point is reached [maximum excursion to the right of the curve representing Eq. (15.21)] at which a few electrons can reach the grid of tube 1. As R_g is large, the potential of this grid then falls rapidly and this is accompanied by a rapid drop in i_{p1} and rise in e_{p1}. This rise in e_{p1}, reacting through C, raises the potential of the grid of tube 2 until this tube conducts, and the cycle is repeated. The interval of conduction of tube 1 is of the order of magnitude of the time constant of its grid circuit. The same remark obviously applies to the second tube so that the period of a complete cycle is of the order of $R_{g1}C_1 + R_{g2}C_2$ if the resistances and capacities are unequal or of the order of $2R_gC$ if they are the same. As can be seen from the figure, the wave that is generated departs widely from a sine curve; it resembles more closely the saw-tooth type produced by the periodic charge and discharge of a condenser (Fig. 7.32). A Fourier analysis of the wave shows that the amplitude of harmonics as high as the hundredth may be quite appreciable. This is very useful for certain purposes. For instance, the output of a crystal oscillator may be loosely coupled (inductively or by means of a small amount of common capacity or resistance) to one of the grid circuits. This will serve to maintain the fundamental (or one of the harmonics) at exactly the crystal frequency. The harmonics are then integral multiples of the crystal frequency and are known with the same degree of precision as the frequency of the crystal itself. This series of accurately known frequencies is very convenient for the calibration of a standard resonant circuit such as a wavemeter.[1]

Problems

1. A type 56 vacuum tube operates in the linear region at a mean plate potential of 250 volts and plate current of 5 ma. What must be the potential of the plate battery for a resistive load of 25,000 ohms? What is then the voltage-amplification constant? ($r_p = 9{,}500$ ohms, $\mu = 13.8$). If an inductance of 10 henrys with a resistance of 100 ohms is available for the plate load instead of the resistance, what battery potential would be necessary? Plot the voltage amplification at 100-cycle intervals from 100 to 1,000 cycles.

[1] ANDREW, *Proc. I.R.E.*, **19**, 1911 (1931).

2. What is the maximum power that the type 56 tube can deliver as a class A amplifier? What is the plate circuit efficiency and what is the maximum alternating-current grid potential required?

3. Find the ratio of the amplitude of the second and third harmonic to the amplitude of the fundamental for a grid swing of 16 volts about $E_c = -16$ volts for a type 56 tube using the load line of Fig. 7.6. What is the change in direct-current on application of the signal? What is the power output at the fundamental and the first two harmonics? Obtain the same data for a grid swing of 8 volts at the same grid bias.

4. Find the second and third harmonic distortion (ratio of the amplitudes of these harmonics to that of the fundamental) for the pentode and load line of Fig. 15.5, using a grid bias of -16 volts and a grid swing of the same magnitude. What is the power output at the fundamental and the change in the mean potential across the load when the signal is applied?

5. Assume that an amplifier having a normal voltage amplification A ($A = \mathbf{e}_{\text{output}}/\mathbf{e}_{\text{input}}$) has the output and input interconnected in such a way that a fraction f of the output potential is returned to the input. (A and f are both real.) Show that the effective or regenerative potential amplification is then $A_r = \dfrac{A(1-f)}{(1-Af)}$ and that if the amplifier requires no power input, $A_r = A/(1 - Af)$.

6. In the amplifier of the preceding problem phase shifts may occur both in the amplifier and in the output-input coupling circuit so that **A** and **f** are in general complex. Assuming infinite input impedance and that $\mathbf{Af} = Af\, e^{j\phi}$, show that

$$A_r = \frac{A}{(1 - 2Af\cos\phi + A^2f^2)^{1/2}}$$

Taking δA_r due to a small change δA as a measure of the stability of the amplifier, show that optimum stability is achieved if $Af\cos\phi = 1$, in which case $A_r = A\cot\phi$.

7. Assuming that the approximate analysis of the class B amplifier is also applicable to the class C type, show that

$$I_{p1} = \frac{\mu_p E_g}{ar_p + R} \qquad \text{where} \qquad a = \frac{\pi}{\theta' - \sin\theta'\cos\theta'}$$

Plot a as a function of θ' at 10° intervals from 10° to 90°.

8. Using Eqs. (15.9) and (15.10), show that the ratio I_{p1}/I_b approaches 2 as θ' approaches 0 leading to a theoretical 100 per cent efficiency in the limit of $E_b = E_{p1}$.

9. Assuming the approximate analysis of the preceding problems and that $E_{p1} = 0.95E_b$, plot the efficiency of a class C amplifier as a function of the half interval of conduction θ' at 10° intervals from 10° to 90°.

10. A parallel-tuned circuit consisting of a variable condenser C and an inductance L of resistance R is connected across the terminals AB of Fig. 15.15. If ω is the observed angular frequency of oscillation, show that

$$L = \frac{1}{2\omega^2C}[1 + (1 - 4R^2C^2\omega^2)^{1/2}] \cong \frac{1}{\omega^2C}(1 - R^2C^2\omega^2)$$

if $2RC\omega$ is small compared to unity. On increasing C to the value C_1, oscillations cease. If a small resistance R' is placed in series with L, oscillations cease at C_2. Show that $R = R'C_2/(C_1 - C_2)$. Show further that the effective negative conductance of the tube is

$$-g = \frac{RC_1}{L} = R\omega^2CC_1(1 + R^2C^2\omega^2)$$

11. Using Eq. (15.15), show that if the load resistance R of a triode satisfies the inequality

$$R\left(R + \frac{r_p C_{gp}}{\mu(C_{gp} + C_{fp}) + C_{gp}}\right) > \left(\frac{\mu}{2\omega[\mu(C_{gp} + C_{fg}) + C_{gp}]}\right)^2$$

instability cannot occur for any value of the load inductance.

12. Using the constants given previously for the type 56 tube and the additional data $C_{gp} = 3.2$ and $C_{fp} = 2.2$, in $\mu\mu$f., find the limits of the load inductance for instability assuming a load resistance of 2,500 ohms.

13. Referring to Fig. 15.19 and assuming that cutoff (θ') occurs at $-\mu e_c = e_b$ show that

$$E_c = \frac{E_b}{\mu} + \left(E_g - \frac{E_b}{\mu}\right)\cos\theta'$$

and hence that the amplitude of grid excitation is given by

$$E_g = \frac{E_b}{\mu} + e_{p\,\min}\frac{\left[\dfrac{\cos\theta'}{\mu} + f\right]}{1 - \cos\theta'}$$

where f is the ratio $e_{p\,\min}/e_{c\,\max}$. Assuming a type 10 tube for which $\mu = 8$, a plate battery of 1,000 volts, an $e_{p\,\min}$ of 100 volts, and an f of 0.8, determine the values of E_g, e_c, and E_p for a θ' of 30°.

14. Design a circuit of type **a** using the negative transconductance of a type 57 pentode to produce a positive reactance depending on the inverse first power of the frequency. Using this circuit, construct one having a reactance depending on the inverse second power of the frequency.

15. Design a circuit of type **b** using the negative transconductance of a type 57 pentode to produce a negative reactance dependent on the first power of the frequency. Using this circuit, construct one in which the reactance depends on the cube of the frequency.

16. A coil with an inductance of 0.1 henry and a resistance of 10 ohms is available for the production of oscillations with a type 57 tube in the negative transconductance circuit. Using the values of the constants given in the text, what is the lowest frequency that can be generated and what is then the value of the capacity?

17. A type 10 tube ($r_p = 5{,}000$ ohms $\mu = 8$) is used in a tuned-plate oscillator circuit with a capacity C of 0.001 μf. and an inductance of 1 millihenry having an effective resistance of 10 ohms. Assuming that the grid coil has an inductance of 0.36 millihenry and that the coefficient of coupling between the two is 0.3, determine the effective plate resistance. Also find the percentage difference between the frequency generated and the natural frequency of the resonant circuit.

18. Neglecting grid current and the inductance of the plate coil, show that the effective plate resistance of a tuned-grid oscillator is

$$r_p = \frac{M}{RC}\left(\mu - \frac{M}{L}\right)$$

and that the frequency of the oscillations is equal to the natural frequency of the resonant circuit.

19. A resistance R and inductance L in series are placed in parallel with a resistance R and capacity C also in series. In series with this combination is a third resistance also of value R and the entire net is placed across the output of an oscillator. If the

resistances have a common junction and $C = 1/(\sqrt{3}\omega R)$ and $L = \sqrt{3}R/\omega$, show that equal potentials $2\pi/3$ out of phase appear across the resistances and hence a three-phase output can be obtained in this way. What is the impedance presented to the oscillator?

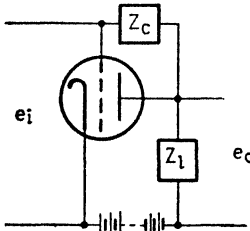
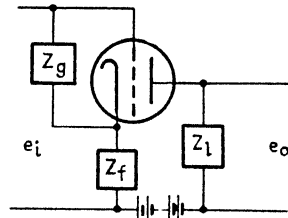
20. How can two-phase output ($\pi/2$ phase difference with equal amplitude) be obtained from an oscillator?

21. Show that the maximum amplitude of oscillation given by Eq. (15.20) is equal to $e_d' - e_c'$ on the basis of the analysis of Fig. 15.12.

22. A symmetrical three-stage resistance-capacity-coupled amplifier has its output returned to the input as in the case of the multivibrator. Assuming that R_p is much less than r_p or R_g, show that to the linear approximation the three circuit equations are

$$D(R_p s_p e_{g1}) + \left(D + \frac{1}{R_g C}\right)e_{g2} + 0 = 0$$
$$0 + D(R_p s_p e_{g2}) + \left(D + \frac{1}{R_g C}\right)e_{g3} = 0$$
$$\left(D + \frac{1}{R_g C}\right)e_{g1} + 0 + D(R_p s_p e_{g3}) = 0$$

where s_p is the transconductance and D stands for the derivative with respect to t. On equating the determinant to zero, show that these equations predict the generation of an undamped sinusoidal wave of period $2\sqrt{3}R_g C$ if $R_p s_p = 2$.

23. In the circuit of the accompanying figure show that $e_0/e_i = (\mu - r_p/z_c)/(1 + r_p/z_c + r_p/z_l)$ and that considered as a feedback amplifier $\beta = \dfrac{(\mu + 1)r_p/z_c + \mu r_p/z_l}{\mu(\mu - r_p/z_c)}$. Plot the frequency characteristic, assuming that z_l is a large resistance and z_c is (a) a resistance, (b) a capacitance, (c) a parallel L-C circuit, (d) a series L-C circuit.

Prob. 23. Prob. 24.

24. In the circuit of the accompanying figure show that $e_0/e_i = \mu z_p z_l/[(z_l + r_p)(z_g + z_f) + \mu z_f z_g]$ and that considered as a feedback amplifier $\beta = (\mu z_f z_g + z_f z_l + z_f r_p + z_g r_p)/(\mu z_l z_g)$. Plot the frequency characteristic, assuming that z_g can be ignored and that z_l is a small resistance for the following types of z_f: (a) resistance, (b) inductance, (c) parallel L-C circuit.

CHAPTER XVI

RADIATION

16.1. Introduction.—As was pointed out in Sec. 10.2, the enlargement of the concept of electric current to bring about consistency between the equations for the divergence of the current and the **curl** of **H** has very important consequences when σ becomes vanishingly small. When the fundamental electromagnetic equations are modified in this way, they predict the possibility of the propagation of an electromagnetic disturbance that has the characteristic spatial and temporal relations associated with progressive wave motion. These waves when in free

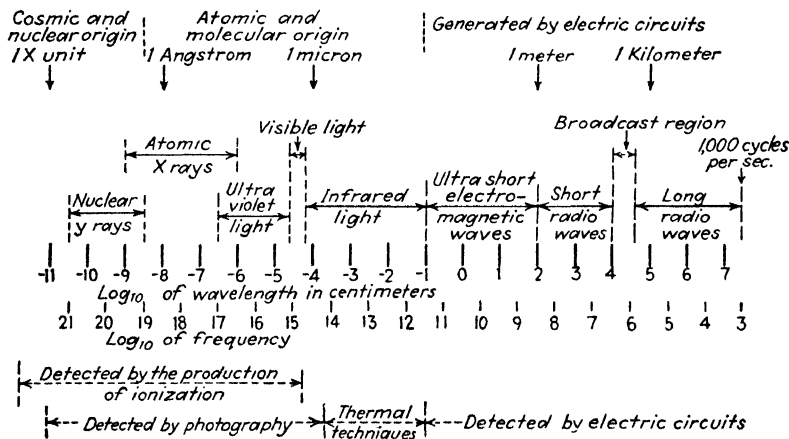



FIG. 16.1.—The electromagnetic spectrum.

space are of the transverse type such, for instance, as would be propagated along an iron bar if it were struck from the side. The prediction of these waves upon Maxwell's assumption of the displacement current was one of the major triumphs of the electromagnetic theory. For it provided a very complete description in electromagnetic terms of all the diverse phenomena associated with the transmission of radiation. These waves occur in nature or can be produced in the laboratory over a very wide range of frequencies or wave lengths. The extent of the electromagnetic spectrum that has been investigated experimentally is indicated in Fig. 16.1. The most familiar region is that of visible light and it is here that the predictions of Maxwell's theory have received their most extensive verification. A complete discussion of the phenomena of

reflexion, refraction, interference, diffraction, polarization, and dispersion is beyond the scope of this book but will be found in any treatise on physical optics. A certain amount of the general theory will be developed in a later section, but more emphasis will be laid on the phenomena associated with those lower frequency electromagnetic waves that are generated by means of electric circuits.

Since these waves traverse free space which is utterly devoid of matter, it was originally felt necessary to introduce an all-pervading medium, the oscillation of which was responsible for the propagation of an electromagnetic disturbance. This is the ether or "luminiferous ether" mentioned in Sec. 2.1. What we observe as an electromagnetic field can be interpreted in terms of a rather complicated system of stresses and strains in this medium. However, little is added to our understanding of electromagnetic phenomena by introducing a medium with the rigidity necessary for the propagation of light but so tenuous that it offers no resistance to the motion of astronomical bodies through it. Furthermore, experiments designed to detect the presence of an ether all lead to results denying its presence.[1] The lack of a medium for the transmission of light is not as serious as it appears at first sight. The formal mathematical description of the phenomena does not depend at all on the presence of a medium. The physical insight gained by introducing an ether would be largely illusory. We are aware of radiation only through the physical effects produced when its energy is absorbed and the physical pulsations that may be attributed to space are not observable and hence are quite irrelevant. A more complete analogy between electromagnetic and mechanical waves would really contribute little to our grasp of the situation for sound waves themselves are without an adequate medium. At one stage in the development of physical theory atoms and molecules could be thought of as small perfectly elastic spheres, and the contact forces between them were considered to be well understood. But as our acquaintance with atomic phenomena has increased, mechanical concepts have lost their usefulness and atomic forces appear to be partially electrical and partially of a type having no large-scale analogue. Thus little would be gained by developing mechanical analogies for electromagnetic waves. Our knowledge of the electrical characteristics of space is as yet very elementary, but a more promising avenue for future development is suggested by Dirac's theory[2] of positive and negative electrons than by any mechanical approach.

16.2. Electromagnetic Waves in Free Space.—In free space there is no net charge density and the conductivity is zero. Thus the four

[1] MICHELSON, "Studies in Optics," University of Chicago Press, 1927. TROUTON and NOBLE, *Proc. Roy. Soc.*, **72**, 132 (1903).

[2] DIRAC, *Proc. Camb. Phil. Soc.*, **26**, 361 (1930).

fundamental differential equations become

$$\text{div } \mathbf{D} = 0 \qquad (2.21)\ (16.1)$$

$$\text{div } \mathbf{B} = 0 \qquad (9.14)\ (16.2)$$

$$\mathbf{curl\ H} = \frac{\partial \mathbf{D}}{\partial t} \qquad (10.7)\ (16.3)$$

$$\mathbf{curl\ E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad (10.3)\ (16.4)$$

Since in free space $\mathbf{B} = \mu_0\mathbf{H}$ and $\mathbf{D} = \kappa_0\mathbf{E}$, on taking the **curl** of Eq. (16.4) one obtains, with the aid of Eq. (16.3)

$$\mathbf{curl\ curl\ E} = -\mu_0\frac{\partial}{\partial t}\mathbf{curl\ H} = -\kappa_0\mu_0\frac{\partial^2\mathbf{E}}{\partial t^2}$$

And in view of Eq. (16.1) the vector identity

$$\mathbf{curl\ curl\ E} = \mathbf{grad}\ \text{div } \mathbf{E} - \nabla^2\mathbf{E}$$

yields

$$\mathbf{curl\ curl\ E} = -\nabla^2\mathbf{E}$$

Hence

$$\nabla^2\mathbf{E} = \kappa_0\mu_0\frac{\partial^2\mathbf{E}}{\partial t^2} \qquad (16.5)$$

Similarly on taking the **curl** of Eq. (16.3) one obtains

$$\nabla^2\mathbf{H} = \kappa_0\mu_0\frac{\partial^2\mathbf{H}}{\partial t^2} \qquad (16.6)$$

These are the characteristic differential equations of wave motion. This type of equation describes a very large group of phenomena, but the present discussion will be limited to the case of an infinite plane wave. By this is meant a wave in which the vector **E** or **H** depends only on one coordinate. If z is chosen for this coordinate, all partial derivatives with respect to x and y vanish. Thus div $\mathbf{E} = 0$ becomes $\partial E_z/\partial z = 0$, hence the electric field has no varying component along the z axis. The vanishing of the divergence of **H** shows that it also has no component along this axis. As this is the direction of propagation of the wave, Eqs. (16.1) and (16.2) show that the wave is of a transverse type, *i.e.*, the electric and magnetic vectors lie in the plane of the wave front. Writing c for the quantity $\{\kappa_0\mu_0\}^{-1/2}$ the component equations of (16.5) become

$$\frac{\partial^2 E_x}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E_x}{\partial t^2} \qquad \frac{\partial^2 E_y}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E_y}{\partial t^2}$$

There will in general be both an x and a y component of the field vector. These components are independent of one another and E_x will be

considered first. If E_x is any function of the form $f\left(t \pm \frac{z}{c}\right)$, the equation is satisfied. This may be verified by substituting the function in the differential equation. Such a function indicates a disturbance propagated along the negative or positive z axis, depending on whether the positive or negative sign is chosen. Thus, if the function has a particular value at the position z_1 and time t_1, it will have the same value at a position z_2 and a later time t_2 if the argument of the function remains unaltered, *i.e.*,

$$t_1 \pm \frac{z_1}{c} = t_2 \pm \frac{z_2}{c} \qquad \text{or} \qquad c = \mp\frac{z_2 - z_1}{t_2 - t_1}$$

Therefore c is the velocity of propagation of the disturbance. (See also Sec. 14.5.) Since κ_0 and μ_0 are numerical constants, this velocity is predicted by the theory to be

$$\begin{aligned} c &= (8.855 \times 10^{-12} \times 1.257 \times 10^{-6})^{-1/2} \\ &= 2.998 \times 10^8 \text{ m. sec.}^{-1} \end{aligned}$$

with an accuracy of approximately one part in 30,000. This prediction has received striking confirmation. The most accurate measurements are those on the velocity of light. This quantity has been determined by many investigators. The most recent work is that of Michelson, Pease, and Pearson.[1] An evacuated path with an equivalent length of 8 or 10 miles was used. The path length and the times involved were measured with extreme accuracy and 2,885 separate determinations were made, leading to a mean value

$$c = (2.99774 \pm 0.00011) \times 10^8 \text{ m. sec.}^{-1}$$

Thus the agreement between the theory and experiment is perfect within the limits of error. This important quantity is one of the most accurately known in all the field of physics.

The relation between the electric field **E** and the associated magnetic vector **H** can be determined from either Eq. (16.3) or Eq. (16.4). In the following discussion the former will be chosen. Since there is only an x component of **E**, there is only an x component of **D**, namely, $\kappa_0 E_x$. Thus there is only an x component of the **curl** of **H** and since the partial derivative with respect to y is zero, this equation becomes

$$-\frac{\partial H_y}{\partial z} = \kappa_0 \frac{\partial E_x}{\partial t}$$

This equation shows that the magnetic vector of the wave is at right angles to the electric one. Choosing a wave traveling in the positive z

[1] MICHELSON, PEASE, and PEARSON, *Astrophys. Jour.*, **82**, 26 (1935).

direction, *i.e.*, $E_x = f\left(t - \frac{z}{c}\right)$, and assuming that the y component of **H** is given by the function $g\left(t - \frac{z}{c}\right)$, the equation becomes

$$\frac{g'}{c} = \kappa_0 f'$$

where the primes indicate differentiation with respect to the argument $\left(t - \frac{z}{c}\right)$. Neglecting any constant fields, which would be unimportant,

$$H_y = \kappa_0 c E_x \tag{16.7}$$

or the more important magnetic induction $B_y = \mu_0 H_y$ is given by

$$B_y = \frac{E_x}{c}$$
$$\cong \frac{E_x}{3 \times 10^8} \text{ webers m.}^{-2}$$

where E_x is, of course, in volts per meter. Thus in an electromagnetic wave in free space the magnetic induction in webers per square meter is less by a factor of 3×10^8 than the electric field strength in volts per meter.

Figure 16.2 illustrates the vector relations in a simple harmonic wave of the form $E_x = E_0 \sin \omega\left(t - \frac{z}{c}\right)$, $H_y = H_0 \sin \omega\left(t - \frac{z}{c}\right)$. The direction of propagation is that of the vector $\mathbf{E} \times \mathbf{H}$. The distance between maxima or any other corresponding points of the wave is evidently the distance that corresponds to a difference of 2π in the argument of the sine function or $z_{m_1} - z_{m_2} = \frac{2\pi c}{\omega}$. This quantity, which is known as the wave length, is generally denoted by λ; in terms of the frequency ν the fundamental relation is $\lambda\nu = c$. A wave, such as that of Fig. 16.2 in which the electric vector (and hence the magnetic vector) is always parallel to one direction, is known as a *plane-polarized wave.* The plane in which the electric vector and the direction of propagation lie will be called the plane of polarization. Such a wave is characteristic of the radiation emitted from a radio antenna; ordinary light can also be rendered plane-polarized. In the more general case the electric vector has a component along the y axis with a corresponding magnetic component in the direction of $-x$. As these components belong to the same wave, they have the same frequency, but they need not be in phase. So in general they would be written $E_x = E_1 \sin \omega t$ and $E_y = E_2 \sin (\omega t + \varphi)$. The trace of the resultant vector in the xy

plane is an ellipse for the most general case, as may be seen by eliminating ωt between these two expressions. The magnitude, eccentricity, and orientation of the major axes depend on E_1, E_2, and φ (see Sec. 7.9). This is known as *elliptically polarized* radiation. Examples may be found in the case of light, or the radiation from a loop carrying a changing current. If E_1 and E_2 are equal and the phase difference is $\pi/2$, $(E_x^2 + E_y^2)^{1/2}$, which is the magnitude of **E**, is a constant and the trace of the extermity of **E** is a circle. This special case is known as *circularly polarized* radiation.

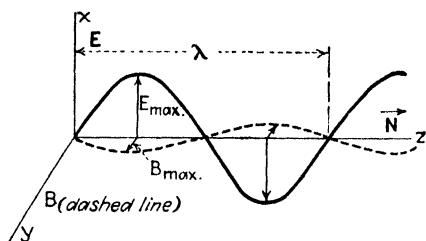



FIG. 16.2.—Vector relations in a simple-harmonic plane-polarized electromagnetic wave.

Our knowledge of radiation comes about through the absorption of the energy that such waves carry. We can get at this quantity most conveniently by multiplying Eq. (16.3) by **E** and subtracting it from Eq. (16.4) multiplied by **H**.

$$\mathbf{H} \cdot \text{curl } \mathbf{E} - \mathbf{E} \cdot \text{curl } \mathbf{H} = -\left(\mathbf{H}\frac{\partial \mathbf{B}}{\partial t} + \mathbf{E}\frac{\partial \mathbf{D}}{\partial t}\right) = -\frac{\partial}{\partial t}\left(\frac{1}{2}\mu_0 H^2 + \frac{1}{2}\kappa_0 E^2\right)$$

The right-hand side is recognized as the rate of decrease of the total electric and magnetic energy density. But by the vector identity

$$\mathbf{H} \cdot \text{curl } \mathbf{E} - \mathbf{E} \cdot \text{curl } \mathbf{H} = \text{div } (\mathbf{E} \times \mathbf{H})$$

this is equal to the divergence of the vector product of **E** and **H**. The integral of this quantity throughout any volume is, by Gauss's theorem, equal to the integral of the normal component of $\mathbf{E} \times \mathbf{H}$ over the bounding surface. The obvious interpretation is that the decrease of energy within the volume is accounted for by an outward flow of energy through the bounding surface equal to the integral of $\mathbf{E} \times \mathbf{H}$ over this surface. The vector representing this energy flow is known as *Poynting's vector* and will be written **N**. From the previous discussion it is seen to be normal to the wave front and in the direction of its motion. Utilizing Eq. (16.7), the rate of flow of energy per unit area due to a plane wave can be written

$$\begin{aligned} \mathbf{N} &= \mathbf{E} \times \mathbf{H} \\ &= \kappa_0 c E^2 \mathbf{n} \end{aligned} \qquad (16.8)$$

where **n** is a unit vector in the direction of propagation of the wave. If **E** is in volts per meter, **N** is equal to $2.65 \times 10^{-3} E^2 \mathbf{n}$ watts per square meter. The absorption of this energy from very short waves produces ionization, effects a photographic plate, or produces the sensation of

sight. This energy can also be absorbed and transformed into heat, and long waves induce the flow of measurable currents in conductors.

When an electromagnetic wave is incident on a conducting or absorbing surface, the theory predicts that it should exert a force on the surface in the direction of **N**. The electric vector **E** is continuous across the boundary of the conductor. Thus a conduction electron is subject to this accelerating force. Its motion in the magnetic field of the wave directs its trajectory in toward the body of the metal and some of its momentum is transferred to the crystal lattice. If q_v is the volume density of charge in the conductor, the body force per unit volume is given by [see Eq. (9.6)]

$$\mathbf{F}_v = q_v(\mathbf{E} + \mathbf{u} \times \mathbf{B})$$

The more general form of the electromagnetic equations taking account of free charges and conduction currents must now be used.

$$\text{div } \mathbf{D} = q_v \qquad \text{and} \qquad q_v\mathbf{u} = \mathbf{curl\ H} - \frac{\partial \mathbf{D}}{\partial t}$$

Substituting these in the force equation and making use of the equality

$$\frac{\partial}{\partial t}(\mathbf{D} \times \mathbf{B}) = \left(\frac{\partial \mathbf{D}}{\partial t}\right) \times \mathbf{B} + \mathbf{D} \times \left(\frac{\partial \mathbf{B}}{\partial t}\right) = \left(\frac{\partial \mathbf{D}}{\partial t}\right) \times \mathbf{B} - \mathbf{D} \times \mathbf{curl\ E}$$

the force per unit volume becomes

$$\mathbf{F}_v = \mathbf{E} \text{ div } \mathbf{D} - \mathbf{B} \times \mathbf{curl\ H} - \mathbf{D} \times \mathbf{curl\ E} - \frac{\partial}{\partial t}(\mathbf{D} \times \mathbf{B})$$

As the mean value of $\mathbf{D} \times \mathbf{B}$ is a constant, and for the plane wave that is being considered $E_y = E_z = H_x = H_z = 0$ and the partial derivatives with respect to x and y also vanish, this equation reduces on the average to

$$\bar{F}_z = -\left(\bar{B}_y \frac{\partial \bar{H}_y}{\partial z} + \bar{D}_x \frac{\partial \bar{E}_x}{\partial z}\right) = -\frac{\partial}{\partial z}\left(\frac{1}{2}\mu_0 \bar{H}_y^2 + \frac{1}{2}\kappa_0 \bar{E}_x^2\right)$$

Thus the average body force on the conductor is equal to the spacial rate of decrease of the average energy density. Considering the boundary layer of thickness dz behind which there is no radiation, the average pressure reaction it must exert is equal to the average energy density of the radiation in front of it. From the previous discussion the average energy density associated with a plane wave which is N/c is equal to $8.85 \times 10^{-12}E^2$ (where E is the effective value of the electric vector). Thus for ordinary values of E of a few volts per meter the pressure is only of the order of 10^{-9} newton per square meter. Though this pressure is extremely small, it was observed by Nichols and Hull[1] using light

[1] NICHOLS and HULL, *Phys. Rev.*, **17**, 26, 91 (1903).

waves. It may also be shown that elliptically or circularly polarized light should exert a torque on being absorbed. This is a still smaller effect but has been observed experimentally by Beth.[1] Of course, if there is no matter to react with the radiation, no pressure will be observed, but the radiation may still be considered to carry momentum with it. By analogy, for instance, with the kinetic theory of gases, the force on a surface may be thought of as the rate of flow of momentum toward it, *i.e.*,

$$\mathbf{F} = \frac{\mathbf{N}}{c} = \mathbf{G}c \qquad \text{or} \qquad \mathbf{G} = \frac{\mathbf{N}}{c^2} = \frac{(\mathbf{E} \times \mathbf{H})}{c^2}$$

where **G** is the effective electromagnetic momentum. This equation is useful in the general discussion of radiation problems, but there will be no occasion to refer to it again in this treatment.

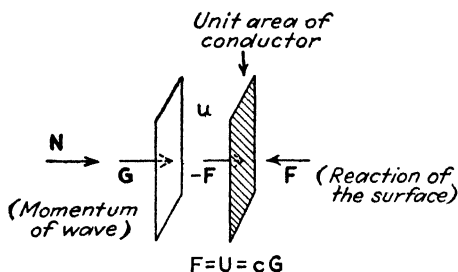



FIG. 16.3.—Mechanical reaction of a surface upon which an electromagnetic wave impinges.

16.3. Extension of the Theory to Include Homogeneous Isotropic Dielectrics.—The extension of the foregoing discussion to take account of the propagation of electromagnetic waves in certain types of material media is very simple. Consider that the substance is a perfect insulator so that the conductivity σ is zero and that it is homogeneous and isotropic, which means that **B** and **D** can be written

$$\mathbf{B} = \mu\mu_0\mathbf{H} \qquad \text{and} \qquad \mathbf{D} = \kappa\kappa_0\mathbf{E}$$

where κ and μ, the dielectric constant and permeability, respectively, are dimensionless constants characteristic of the medium. The fundamental equations are then the same as (16.1) to (16.4), but in the subsequent development $\kappa_0\mu_0$ becomes $\kappa\mu\kappa_0\mu_0$. Therefore the velocity of propagation of the wave c' is $(\kappa\mu\kappa_0\mu_0)^{-1/2}$ instead of $(\kappa_0\mu_0)^{-1/2}$ or

$$c' = \frac{c}{\sqrt{\kappa\mu}} \tag{16.9}$$

where c is the velocity of the wave in free space. Except for ferromagnetic substances, which are unimportant for this discussion, μ is

[1] BETH, *Phys. Rev.*, **50**, 115 (1936).

very close to unity and $c' = c/\sqrt{\kappa}$. Since κ is always greater than 1, the velocity of propagation of a wave in a dielectric is less than its velocity in free space. It is customary to discuss the properties of a dielectric in terms of its *index of refraction*, n, which is defined as the ratio c/c'. Thus $n = \sqrt{\kappa}$. The quantities n or c' can be measured in various ways for radiation and the experimental values for low frequencies are in excellent agreement with the predictions of the theory using values of κ determined electrostatically. At higher frequencies, corresponding to visible light and beyond, the agreement becomes unsatisfactory and the simple theory developed here is shown to be inadequate. In this region it is necessary to take account of the absorption and re-emission of the wave by atoms in its path; these considerations lead to the phe-

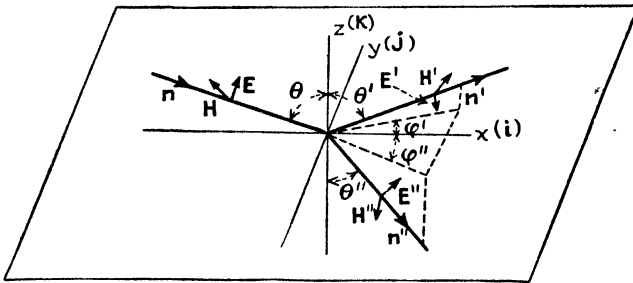



FIG. 16.4.—Schematic representation of a wave normal or ray at the boundary between two dielectric media.

nomena of dispersion and absorption which are beyond the scope of this treatment.

A great deal of interesting and useful information comes from the analysis of the behavior of a wave at a boundary between two dielectric media. In general, the wave normal will have components along all three Cartesian axes and can be written in vectorial form

$$\mathbf{E} = \mathbf{E}_0 \sin \omega\left[t - \frac{(\mathbf{n} \cdot \mathbf{r})}{c'}\right] \tag{16.10}$$

Here $\mathbf{n}$ is a unit vector in the direction of the wave normal and $\mathbf{r}$ is the radius vector $\mathbf{i}x + \mathbf{j}y + \mathbf{k}z$. It will be assumed that the wave length is large in comparison with the thickness of the interface which will be taken as the plane $z = 0$. In general there will be a reflected wave $\mathbf{n}'$ lying in the same medium as $\mathbf{n}$ and a refracted wave $\mathbf{n}''$ in the other medium. This situation is illustrated in Fig. 16.4. To allow for a possible phase shift in these two waves, they would be written

$$\mathbf{E}' = \mathbf{E}_0' \sin\left\{\omega'\left[t - \frac{(\mathbf{n}' \cdot \mathbf{r})}{c'}\right] - \delta'\right\}$$

and

$$\mathbf{E}'' = \mathbf{E}_0'' \sin \left\{ \omega'' \left[t - \frac{(\mathbf{n}'' \cdot \mathbf{r})}{c''} \right] - \delta'' \right\}$$

where c'' is the velocity of propagation in the second medium and the δ's are the phase shifts. The electric and magnetic conditions at the boundary are that the tangential components of **E** and **H** shall be continuous. These conditions may be written

$$\begin{aligned} E_x + E_x' &= E_x'' \qquad & H_x + H_x' &= H_x'' \\ E_y + E_y' &= E_y'' \qquad & H_y + H_y' &= H_y'' \end{aligned} \tag{16.11}$$

In order to apply these, **n**, **n**$'$, and **n**$''$ must be written in terms of their components. Choosing **n** in the xz plane and the angles shown in Fig. 16.4, unit vectors in the directions of the wave normals are

$$\begin{aligned} \mathbf{n} &= \mathbf{i} \sin \theta - \mathbf{k} \cos \theta \\ \mathbf{n}' &= \mathbf{i} \sin \theta' \cos \phi' + \mathbf{j} \sin \theta' \sin \phi' + \mathbf{k} \cos \theta' \\ \mathbf{n}'' &= \mathbf{i} \sin \theta'' \cos \phi'' + \mathbf{j} \sin \theta'' \sin \phi'' - \mathbf{k} \cos \theta'' \end{aligned}$$

For Eqs. (16.11) to be satisfied at all times and at all points on the surface the arguments of the sine functions must all be effectively the same. This means first that the frequency is unaltered or $\omega'' = \omega' = \omega$. Likewise the δ's are equal to either zero or π (the latter simply altering the sign of the amplitude). A further consequence is that since **n** contains no y component, this component must vanish for **n**$'$ and **n**$''$, or $\phi' = \phi'' = 0$. Thus all the ray vectors lie in the plane perpendicular to the surface that contains the incident ray. For equality of the x component of **n** and **n**$'$ it is necessary that $\theta = \theta'$. That is, the angle between the incident ray and the normal to the surface is the same as that between the normal and the reflected ray. This is the fundamental law of reflection. The necessary condition on this component of **n**$''$ is that

$$\frac{\sin \theta}{c'} = \frac{\sin \theta''}{c''}$$

or

$$n' \sin \theta = n'' \sin \theta'' \tag{6.12}$$

where n' and n'' are the indices of refraction of the two media. This is the fundamental law of refraction and is known as *Snell's law*. If n'' is greater than n' (as from air into glass), θ must be greater than θ'' and the ray is bent toward the normal. A refracted ray exists for all possible values of θ in this case. But if n' is greater than n'' (as from glass into air), the maximum value of $\sin \theta$ for which a refracted ray will

exist is given by n''/n' for then $\sin \theta''$ has its greatest possible value, which is unity, and for still larger values of θ no ray will enter the medium of smaller index. This limiting value of θ is known as the *critical angle* for total reflection.

The amplitudes of the electric and magnetic vectors in the reflected and refracted rays can be determined from Eqs. (16.11). For this purpose it is more convenient to consider the electric vector as the sum of two components, E_i lying in the plane of incidence and E_s lying in the plane of the bounding surface. On referring to Fig. 16.4, the first of Eqs. (16.11) is seen to yield

$$E_i \cos \theta - E_i' \cos \theta = E_i'' \cos \theta''$$

and the second

$$E_s + E_s' = E_s''$$

When $\kappa\kappa_0$ and $\mu\mu_0$ are substituted for κ_0 and μ_0 to take account of the properties of the dielectric, Eq. (16.7) becomes

$$H_y = \kappa\kappa_0 c' E_x = \sqrt{\frac{\kappa\kappa_0}{\mu\mu_0}} E_x$$

Since μ is taken as approximately unity and $\sqrt{\kappa}$ is equal to the index of refraction n, this can be written in general vector form as

$$\mathbf{H} = n\sqrt{\frac{\kappa_0}{\mu_0}} (\mathbf{n} \times \mathbf{E}) \tag{16.13}$$

Thus H_x is proportional to nE_y and H_y is proportional to nE_x and the second two of Eqs. (16.11) yield

$$n'(E_s - E_s') \cos \theta = n'' E_s'' \cos \theta''$$
$$n'(E_i + E_i') = n'' E_i''$$

These four equations are sufficient to determine the four unknowns E_i', E_s', E_i'', and E_s''. After eliminating n' and n'' by means of Eq. (16.12), these are obtained explicitly as

$$E_i' = E_i \frac{\tan (\theta - \theta'')}{\tan (\theta + \theta'')}$$
$$E_s' = -E_s \frac{\sin (\theta - \theta'')}{\sin (\theta + \theta'')}$$
$$E_i'' = E_i \frac{2 \sin \theta'' \cos \theta}{\sin (\theta + \theta'') \cos (\theta - \theta'')}$$
$$E_s'' = E_s \frac{2 \sin \theta'' \cos \theta}{\sin (\theta + \theta'')}$$

These are known as *Fresnel's equations.* Through Eqs. (16.8) and (16.13) they determine the energy in the reflected and refracted beams. Like-

wise as the angle between the incident plane of polarization and the xz plane is $\tan^{-1} (E_i/E_s)$, the tangents of the analogous angles for the reflected and refracted beams are (E_i'/E_s') and (E_i''/E_s''), respectively. The first of Fresnel's equations has a particularly interesting consequence. For a certain angle of incidence it is possible to have the refracted ray perpendicular to the reflected ray, *i.e.*, $\theta + \theta'' = \pi/2$. In this case the tangent in the denominator is infinite and hence E_i' is zero. Thus no light reflected at this particular angle has an electric component in the plane of incidence. Using Eq. (16.12), this angle of incidence, which is known as the polarizing angle θ_p, is seen to be

$$\tan \theta_p = \frac{n''}{n'} = n$$

This expression is known as *Brewster's law*. If ordinary light with random planes of polarization is incident on the surface at this angle, the reflected light will be completely polarized with its electric vector normal to the plane of incidence. This is one of the standard methods of obtaining plane-polarized light.

16.4. Propagation of Light in a Conducting Medium.—In order to discuss electromagnetic radiation in a medium that contains free charges it is necessary to take account of the conduction current. That is, the complete equation for **curl H** [Eq. (10.7)] must be used instead of the simplified Eq. (16.3). Assuming that the medium is homogeneous and isotropic so that the conduction-current density $\mathbf{i}_v$ can be written as $\sigma\mathbf{E}$, where the conductivity, σ, is a constant this equation becomes:

$$\mathbf{curl\ H} = \sigma\mathbf{E} + \frac{\partial \mathbf{D}}{\partial t}$$

It will be further assumed, as in the previous sections, that κ and μ are constants. On taking the **curl** of Eq. (16.4) and substituting the complete expression for the **curl** of **H,** the following equation is obtained

$$\mathbf{curl\ curl\ E} = -\mu\mu_0\frac{\partial}{\partial t}(\mathbf{curl\ H}) = -\mu\mu_0\left(\sigma\frac{\partial \mathbf{E}}{\partial t} + \kappa\kappa_0\frac{\partial^2 \mathbf{E}}{\partial t^2}\right)$$

Since Eq. (16.1) still applies, **curl curl E** $= \nabla^2\mathbf{E}$ and writing

$$c' = (\kappa\kappa_0\mu\mu_0)^{-\frac{1}{2}},$$

the equation becomes

$$\nabla^2\mathbf{E} = \mu\mu_0\sigma\frac{\partial \mathbf{E}}{\partial t} + \frac{1}{c'^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{6.14}$$

This is the complete equation for the vector **E** in a conducting medium. An exactly similar equation can be obtained for **H.** The subsequent

discussion of this equation will be limited to the case of a plane wave traveling in the direction of the z axis with its electric vector in the direction of the x axis. The equation for E then reduces to

$$c'^2\frac{\partial^2 E_x}{\partial z^2} = \frac{\sigma}{\kappa\kappa_0}\frac{\partial E_x}{\partial t} + \frac{\partial^2 E_x}{\partial t^2} \tag{16.14'}$$

It is more convenient to discuss the solution of this differential equation in terms of the exponential rather than the sine function. Therefore it will be understood that E_x is either the real or the imaginary part of the following expression:

$$E_x = E_0 e^{j\omega\left(t - \frac{z}{v'}\right)} \tag{16.15}$$

On substituting this in Eq. (16.4) and recalling that E_y and E_z are zero and that the partial derivatives with respect to x and y vanish, it is found that

$$\frac{\partial E_x}{\partial z} = -\frac{\partial B_y}{\partial t}$$

or

$$B_y = \frac{E_x}{v'}$$

This is similar to the result obtained in Sec. 16.2, which shows that the relation between the electric and magnetic vectors is formally the same for a conducting and for a nonconducting medium.

To verify that Eq. (16.15) is a solution it must be substituted in Eq. (16.14'). On performing this substitution and dividing through by E_x the following equation is obtained which must be satisfied by v' in order that Eq. (16.15) may be a solution:

$$\left(\frac{c'}{v'}\right)^2 = 1 - \frac{j\sigma}{\omega\kappa\kappa_0} \tag{16.16}$$

This shows that in general v' is complex, which implies a phase difference between the magnetic and electric vectors. In fact, unity can be neglected in comparison with the purely imaginary part if

$$\sigma >> \omega\kappa\kappa_0. \qquad \text{(See Sec. 10.2)}$$

Metallic conductivities are of the order of 10^7 mhos per meter, therefore unity can be neglected if ω is less than 10^{16}. As this corresponds to ultraviolet light, it is evident that $(c'/v')^2$ can be taken as a pure imaginary for ordinary electromagnetic waves in a metal. Extracting the square root of Eq. (16.16), neglecting unity, it is found that

$$\frac{1}{v'} = \frac{\pm j(1 + j)}{\omega\delta} \tag{16.17}$$

where $\delta = (2/\mu\mu_0\omega\sigma)^{1/2}$. Choosing the negative sign to represent propagation in the direction of positive z and substituting for v' in Eq. (16.15)

$$E_x = E_0 e^{-(z/\delta)} e^{j\omega(t - z/\omega\delta)}$$

The last exponential represents a disturbance propagated with a velocity $\omega\delta$ and the second factor indicates an exponential damping as the wave progresses into the metal. The distance δ is known as the *skin depth*. E_0 is the electric intensity at the surface of the metal which is considered to be the xy plane. For small values of δ the velocity of propagation in the metal is small and hence the wave length of the radiation is smaller than outside the surface.

It is more instructive to consider the damping as a function of the distance in from the surface in terms of the vacuum wave length $\lambda_0 = 2\pi c/\omega$. In terms of λ_0, δ is given by

$$\delta = \left(\frac{\lambda_0\sqrt{\kappa_0/\mu_0}}{\pi\mu\sigma}\right)^{1/2} \tag{16.18}$$

E_x will be reduced to $1/e$ or 0.368 of its value at the surface after it has traversed a distance δ in the metal. To take the instance of copper for which $\mu = 1$ and $\sigma = 5.80 \times 10^7$ mhos per meter, $\delta = 3.82 \times 10^{-6}\sqrt{\lambda_0}$, the distance corresponding to a decrease to approximately a thousandth of its surface value is $2.64 \times 10^{-5}\sqrt{\lambda_0}$m. Thus a 0.1-mm. wave is reduced to this value in about $\frac{1}{4}$ micron, a wave 1 meter long penetrates less than 0.01 mm., and even a 100-m. wave in the broadcast region has an effective value for only about $\frac{1}{4}$ mm. The general behavior of the electric vector is indicated schematically in Fig. 16.5, though no attempt has been made to draw the wave inside the surface to scale. Since the magnetic vector is inversely proportional to v', it is inversely proportional to δ. As δ is small, the energy of the wave inside the metal is practically all in the magnetic form. As δ is a measure of the effective penetration, a wave will penetrate much farther into a poor conductor. On the other hand, for a ferromagnetic material such as iron, though σ is of the order of $\frac{1}{6}$ that for copper, μ may be 600 times as large, which means that the penetration of a wave in iron is only about $\frac{1}{10}$ of its penetration in copper.

The preceding equations describe the behavior of the electromagnetic wave in a metal, but it is of importance as well to relate this wave to the one in the free space outside the plane metal surface. This can be done by extending the argument of Sec. 16.3. For simplicity the discussion will be limited to normal incidence, as this illustrates the features of particular interest, although the change in polarization that is brought about at oblique incidence is important in determining the optical properties of metals. As the electric and magnetic vectors are parallel to the boundary, the conditions on the continuity of **E** and **H** there

become

$$\mathbf{E}_i + \mathbf{E}_r = \mathbf{E}_t, \qquad \mathbf{H}_i + \mathbf{H}_r = \mathbf{H}_t$$

The relations between **E** and **H** in free space and in the metal are, respectively,

$$\mathbf{H}_{\text{space}} = \frac{\mathbf{n} \times \mathbf{E}_{\text{space}}}{\mu_0 c}, \qquad \mathbf{H}_{\text{metal}} = \frac{\mathbf{n} \times \mathbf{E}_{\text{metal}}}{\mu_0 v'}$$

where **n** is a unit vector in the direction of propagation ($\mathbf{n}_i = \mathbf{n}_t = -\mathbf{n}_r$).

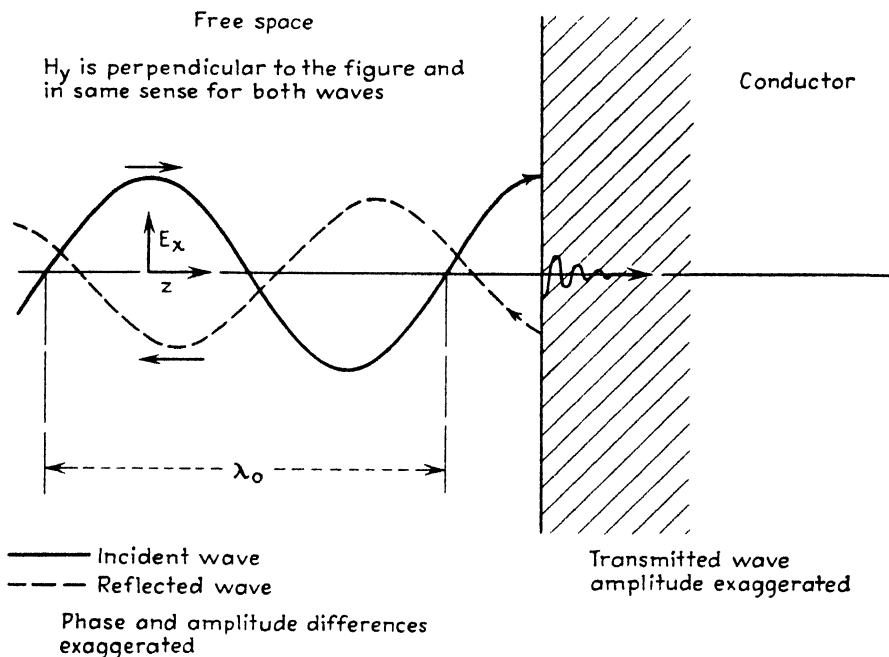



FIG. 16.5.—Schematic representation of the reduction in amplitude of an electromagnetic wave on entering a metal.

Taking the vector product of $\mathbf{n}_i$ and the equation for the **H**'s the boundary conditions become

$$\mathbf{E}_i + \mathbf{E}_r = \mathbf{E}_t, \qquad \mathbf{E}_i - \mathbf{E}_r = \frac{c}{v'}\mathbf{E}_t$$

or as v'/c is very small in absolute value for metals,

$$\mathbf{E}_t \cong \frac{2v'}{c}\mathbf{E}_i, \qquad \mathbf{E}_r \cong -\left(1 - \frac{2v'}{c}\right)\mathbf{E}_i$$

where only the first powers in v'/c have been retained. The ratios E_t/E_i and E_r/E_i are known as the *transmission* and *reflection coefficients*, respectively, and the analogy with the Γ's of transmission-line theory is seen to be quite close (Sec. 14.5). The fraction of energy reflected is proportional

to $|E_r/E_i|^2$. Writing this as R_r, and $1 - R_r$ as R_t, which is the fraction of the energy absorbed, these ratios are seen to be

$$R_r \cong 1 - \frac{2\omega\delta}{c}, \qquad R_t \cong \frac{2\omega\delta}{c} \tag{16.19}$$

where v' has been taken from Eq. (16.17). By far the greater fraction of energy is reflected from the metal, only about 1 per cent of it being absorbed by copper for a wave length of 1 cm.

It is somewhat more convenient for calculation to have the flow of energy into the metal given in terms of the total field vectors at the surface. Since $\mathbf{E}_r \cong -\mathbf{E}_i$ and $n_r = -n_i$, the equation for $\mathbf{H}$ in the space in front of the metal shows that $\mathbf{H}_r \cong \mathbf{H}_i$ or the total surface component of $\mathbf{H}_t$ is $2\mathbf{H}_i$ to the zeroth order of small quantities. The energy transported in the incident wave per unit area per unit time is $\mathbf{N}_i = \mathbf{E}_i \times \mathbf{H}_i = \mu_0 c\mathbf{H}_i^2\mathbf{n}$, and from Eq. (16.19) the rate at which energy enters the metal per unit area is

$$\frac{dU}{dt} = \frac{2\omega\delta}{c} \cdot \mu_0 c H_i^2 \cong \frac{\omega\delta}{2}\mu_0 H_t^2 \tag{16.19'}$$

The above assumes a plane surface but is a satisfactory approximation if the radius of curvature of the surface is large compared with the skin depth δ. This is the basic relation for determining the dissipation of electromagnetic waves at the surfaces of conductors.

Propagation through Ionized Regions.—The propagation of long electromagnetic waves, in the radio range, through a region containing free ions and electrons is of interest for certain of the phenomena observed are of great practical importance. It is known from several lines of evidence that an appreciable number of the atoms and molecules composing the earth's atmosphere are ionized at a height of several hundred kilometers above the surface. This blanket of ionization is known as the ionosphere or the *Kennelly-Heaviside layer.*[1] It is not a permanent well-defined region but varies in effective height and ion density from hour to hour. The ion density as a function of the height undoubtedly depends on many factors, most of which are not well understood. However, the major portion of the ionization is doubtless produced by the various types of solar radiation. The ionization is much greater in the day time than at night and it shows a marked dependence on sunspot activity and so-called magnetic storms. There is a tendency for the ion density to occur in strata or layers which are constantly fluctuating in height and character. Though the phenomena associated with these layers are being intensively studied, it is not as yet possible to account adequately

[1] For a more detailed discussion of the ionized layers see: APPLETON, *Inst. E. E.*, **7**, September, 1932; KIRBY, BERKNER, and STUART, *Proc. I.R.E.*, **22** (1934); MIMNO, *Rev. Mod. Phys.*, **9**, 1 (1937); DARROW, *Bell. System Tech. J.*, **19**, 455 (1940).

for their behavior. It is known, however, that they are responsible for the long-range propagation of radio waves over the earth's surface, and the cause of this will be briefly considered.

If E_x is the electric vector of the wave, the electric force on an ion of charge e that is encountered is given by eE_x. The effect of the magnetic vector associated with the wave is smaller by the factor $1/c$ and can be neglected. Therefore if the ion has a mass m, the equation of motion is

$$m\frac{d^2x}{dt^2} = eE_x$$

Assuming a simple harmonic wave of the form of Eq. (16.15), this differential equation can be integrated directly to give the displacement as

$$x = -\frac{e}{m\omega^2}E_x$$

Thus the displacement is in the opposite direction to the electric force and is inversely proportional to ω^2. Damping or dissipative forces have been neglected, but it is reasonable to assume that energy will be lost from the wave only through collisions of the ion with molecules in its path. At such a collision the ion can lose some of its kinetic energy and this loss of energy will be proportional to the path length. Thus, for small values of ω, x is great and energy is rapidly lost, but for high frequencies the damping is small and to a first approximation can be neglected. If the ion density per cubic meter is N, the conduction current due to the freely moving ions is given by

$$i_x = Ne\frac{dx}{dt} = -\frac{je^2N}{\omega m}E_x$$

or the conductivity σ of the medium which is the ratio i_x/E_x is

$$\sigma = -\frac{je^2N}{\omega m} \tag{16.20}$$

This conductivity is seen to be a pure imaginary quantity which means that the current and field waves are $\pi/2$ out of phase with one another and to this approximation no power is consumed. Though the wave is not damped, the velocity of propagation is altered by the presence of the ions. Inserting Eq. (16.20) in Eq. (16.16) and taking κ as unity, the index of refraction is seen to be

$$n = \frac{c}{v'} = \left(1 - \frac{Ne^2}{m\kappa_0\omega^2}\right)^{1/2} \tag{16.21}$$

The second term in the parentheses is positive for either positive ions or electrons so the negative sign indicates that the velocity of propagation

of the wave through the ions is greater than its velocity in free space. This does not violate the general principle of relativity, which states that no signal can be propagated with a velocity greater than that of light in free space, for such a signal implies the existence of a group of waves of different wave lengths and it can be shown that the velocity v_g with which such a group is propagated, is given by $v_g = c^2/v'$. Hence the group velocity is less than c. As v' is larger than c and real the magnetic vector is smaller than it would be in free space and it is in phase with the electric vector.

Consider a plane wave sent off from the surface of the earth that encounters a region in which the ion density increases with the height. The upper portion of the wave travels more rapidly than the lower portion and the wave is refracted back toward the surface of the earth. This is indicated schematically in Fig. 16.6. From the construction it is evident that

$$\frac{dv'}{dh} = \frac{v'}{\rho}$$

where ρ is the radius of the curved wave at a point where its velocity is v'. From this it is evident that if the ion density increases with h, the wave is refracted back toward the earth, but if the ion density decreases, the wave is bent away from the earth and lost. It is essentially a refraction problem and Eq. 16.12 in the general form $n \sin \theta =$ const. can be applied. Let θ be the angle between the wave normal and the normal to the surface of the earth as the wave approaches the layer. In order that the wave shall reach a maximum height at which it is traveling parallel to the earth's surface, θ'' at this height must be equal to $\pi/2$. Assuming that n' beneath the layer is unity n'' must evidently equal $\sin \theta$. On writing n of Eq. (16.21) for n''

FIG. 16.6.—Refraction of a wave on encountering a region of increasing ion density.

$$\sin \theta = \left(1 - \frac{Ne^2}{m\kappa_0\omega^2}\right)^{1/2}$$

or

$$\cos \theta = \left(\frac{Ne^2}{m\kappa_0\omega^2}\right)^{1/2} \tag{16.22}$$

If ω and θ are known, the corresponding ion density N at the maximum height of the wave can be calculated. The order of magnitude of the maximum value of N can be obtained from the highest frequency for

which reflection is observed under the most favorable conditions (maximum θ). This is approximately 3×10^7 cycles per second in the day time. In order to obtain values of N that are reasonably in accord with other experimental evidence, it must be assumed that the charged particles principally responsible for the phenomenon are electrons. On inserting the appropriate numerical values the right-hand side of Eq. (16.22) becomes $2.8 \times 10^{-7}N^{1/2}$. The most favorable case for observation is the one involving the largest value of θ for which the wave is reflected back to the earth. This path is indicated by 3 of Fig. 16.7 and both leaves and approaches the earth's surface tangentially. Taking the mean radius of the earth as 6.37×10^6 m. and the height of the ionized layer as approximately 3×10^5 m., a simple geometric construction

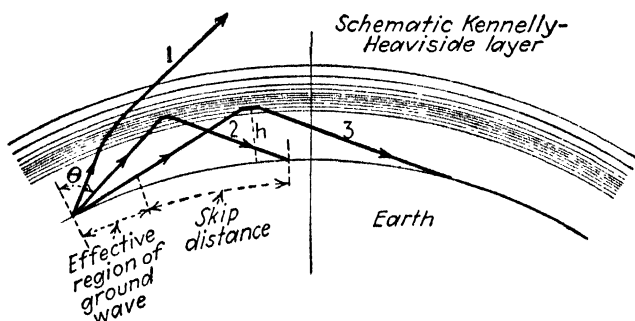



FIG. 16.7.— Refraction of radio waves from the ionosphere.

shows that the largest possible angle between the wave and the normal to the layer is about 75°. This corresponds to a cos θ of 0.26, and Eq. (16.22) gives the maximum value of N as 8.5×10^5 free electrons per cubic centimeter.

Figure 16.7 illustrates schematically the characteristic behavior of the ionized layer in refracting a radio wave back to the surface of the earth. The so-called *ground wave* is propagated more or less directly from the transmitter to the receiver without reflection from the ion layer, though ionization at low altitudes undoubtedly plays a role in deflecting the wave fronts toward the earth. Its distance of effectiveness is rather limited, as indicated in Figs. 16.7 and 16.8. The point at which the sky wave first returns to the earth is determined by the angle at which the wave leaves the earth (effectively the angle at which the radiated energy is a maximum) and the height and constitution of the ionized layer. In general, there is a region beyond the limit of the ground wave, but too close for reception of the sky wave. This is known as the skip distance. Waves have been observed to travel around the earth several times, giving rise to "radio echoes." This is most probably accomplished by multiple reflection from the ionized layers and the earth's surface. If one is situated near the limit of return of the sky wave or where

interference effects between the ground wave and the sky wave are observable, the fluctuations of density and height of the layers will give rise to large variations of the received signal intensity. This is known as "fading." Figure 16.8 indicates the approximate limits for the propagation of radio waves under various conditions. The curves are to be regarded as approximate only. The behavior of any one wave length is determined by following a horizontal line across the diagram; the signal is detectable between the appropriate pair of curves. It is

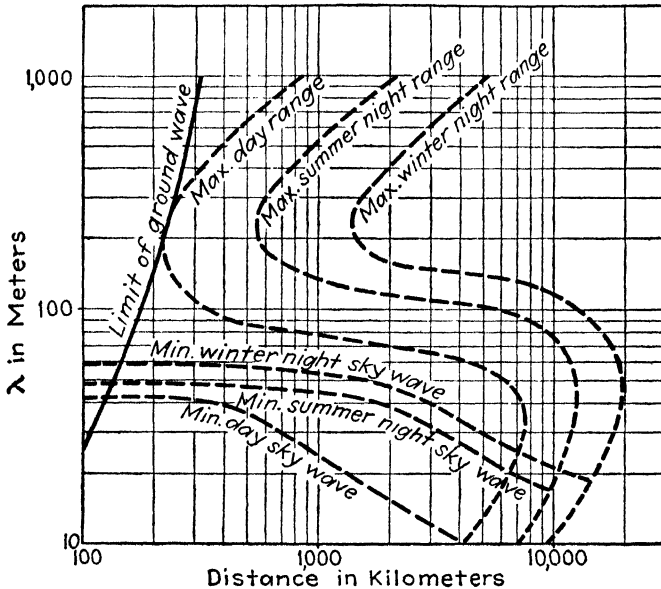



FIG. 16.8.—Propagation of radio waves. (*After A. H. Taylor "Radio Amateur's Handbook."*)

assumed that 5 watts are radiated by the antenna and that the minimum detectable signal is 10 μv. per meter.

The nature of the propagation of an electromagnetic wave through a region in which the magnetic induction is **B** and there are N electrons per unit volume presents the interesting feature that different indices of refraction are associated with the two senses of circular polarization (Sec. 16.2). Assume that the induction is in the z direction, and write the equations of motion from Eq. (9.6)

$$\frac{m\,du_x}{dt} = e(E_x + u_y B_z)$$

$$\frac{m\,du_y}{dt} = e(E_y - u_x B_z)$$

$$\frac{m\,du_z}{dt} = eE_z$$

Here e and m are the charge and mass, respectively, of the electron and u is its velocity. The equation for the z component presents nothing new, so it will be assumed that z

is the direction of propagation of the wave having electric components E_x and E_y. The effect of the magnetic vector associated with the wave is less than that of E by the factor $1/c$ and hence is neglected. The effect of any energy loss due to collisions with gas molecules is neglected as well. By multiplying the y equation by j and adding to the x equation it is seen that the single equation for $u = u_x + ju_y$ results:

$$\frac{m\,du}{dt} = e(E - juB)$$

where $E = E_x + jE_y$ and B, which is perpendicular to E, and u is written for B_z. This equation and Eq. (16.14), in which Neu is written for σE, determine the nature of the propagation. Assuming solutions in the form of right and left circularly polarized

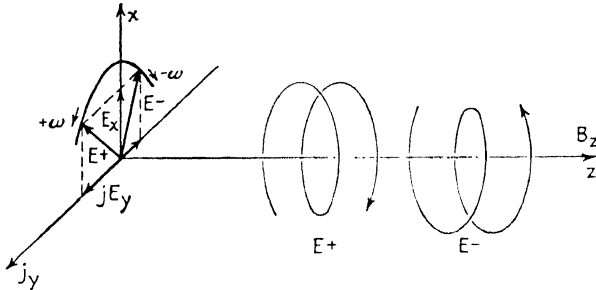



FIG. 16.9.—Propagation of an electromagnetic wave through a region of magnetic induction containing free charges.

waves traveling in the z direction, $E = E_0e^{\pm j\omega(t-nz/c)}$ and $u = u_0e^{\pm j\omega(t-nz/c)}$; and taking $\kappa = \mu = 1$, the equations for the magnitudes of u and E become

$$n^2E = E \mp j\frac{\mu_0 Nec^2}{\omega}$$
$$\pm jm\omega u = e(E - juB)$$

Eliminating u and E, the indices of refraction corresponding to the choice of signs are seen to be given by

$$n_{\pm}^2 = 1 - \frac{Ne^2}{\kappa_0\omega(\omega m \pm Be)}$$

The vector E_+ of Fig. (16.9) is associated with the positive sense of rotation ($e^{j\omega t}$), and the vector E_- with the negative sense ($e^{-j\omega t}$). Although the numerator of the second term is generally small, the denominator may become very small for E_- when ω is close to eB/m. The damping factors that prevent this term from becoming infinite have been neglected, and it is clear that even if they are included, the propagation of this mode will be anomalous near this critical frequency. It will be noted that this is the characteristic resonant frequency discussed in Sec. 9.2, and in its neighborhood the amplitude of motion of the electrons becomes very great. This results in a much more rapid loss of energy to the molecules in its path and a large damping for a wave of this particular frequency. The average effective value of the earth's magnetic field in the ionosphere is approximately 0.5 gauss or 5×10^{-5} weber per square meter, corresponding to a resonant frequency of about 1.4×10^6 cycles or a wave length of 215 m. This is very close to the wave length for which maximum absorption occurs in the ionosphere, as shown in Fig. 16.8. The numerical agreement with the observed absorption maximum is more or less fortuitous as the effective value

of B varies with the direction of propagation relative to the earth's field. This variation is mainly responsible for the observed breadth of the absorption curve. The equation for $n_{\pm}^2$ accounts in large measure for the propagation phenomena exhibited by radio waves traversing the ionized regions of the upper atmosphere and in particular for the elliptical polarization of the wave returned to the earth. The fluctuations of the ionized region alter the nature of this polarization, and this further contributes to the fading of radio waves.

16.5. Propagation of Electromagnetic Waves in Metallic Enclosures. The conditions that are imposed on electromagnetic waves by the conducting boundaries of a region result in certain characteristic phenomena of great practical importance. The propagation of electromagnetic waves in metal pipes was first studied by Rayleigh,[1] and the properties of electromagnetic waves in cavities came into prominence in connection with the role of black-body radiation in the early development of the quantum theory.[2] More recently it has assumed much greater importance in connection with radar and high-frequency communication circuits. The conditions imposed by the conducting boundaries introduce relations between the wave lengths and directions of propagation that limit the modes of oscillation that can be sustained. For simplicity our consideration will be limited to rectangular enclosures or cavities in conductors, but the general nature of the phenomena to be described is independent of size or shape. It will further be assumed that the cavities contain no dielectric or magnetic material; *i.e.*, $\kappa = \mu = 1$.

Consider first a rectangular metal box bounded by the planes $x = 0$ and A, $y = 0$ and B, $z = 0$ and C (Fig. 16.10). If a solution for Eq. (16.5) is assumed of the form $\mathbf{E} = \mathbf{E}'e^{\pm j\omega t}$, this equation becomes

$$\nabla^2\mathbf{E}' + k^2\mathbf{E}' = 0 \tag{16.23}$$

where $k^2 = (\omega/c)^2$. The analogous equation exists for the accompanying vector $\mathbf{H}'$ and the relation between $\mathbf{E}'$ and $\mathbf{H}'$ given by Eq. (16.4) as $\mathbf{H}' = \pm(j/\mu_0\omega)$ **curl** $\mathbf{E}'$. A vector solution of Eq. (16.23) can be written in the form

$$\mathbf{E}' = \mathbf{E}_0e^{\pm j\mathbf{k}\cdot\mathbf{r}}$$

where $\mathbf{k}$ is the so-called *propagation vector* $\mathbf{i}k_x + \mathbf{j}k_y + \boldsymbol{\kappa}k_z$, as may be seen by substitution. ($\boldsymbol{\kappa}$ is a unit vector along z.) Thus the general plane wave propagated in the direction $\mathbf{k}$ may be written

$$\mathbf{E} = \mathbf{E}_0e^{\pm(j\omega t - \mathbf{k}\cdot\mathbf{r})}$$

The metal boundaries can be considered so nearly perfectly conducting that the boundary conditions may be written

[1] RAYLEIGH, *Phil. Mag.*, **43**, 125 (1897).

[2] JEANS, "The Dynamical Theory of Gases," 3d ed., Cambridge University Press, London, 1921.

$$E_y = E_z = 0 \text{ at } x = 0 \text{ or } A, \qquad E_x = E_z = 0 \text{ at } y = 0 \text{ or } B, \qquad E_x = E_y = 0 \text{ at } z = 0 \text{ or } C$$

The electric field in the cavity must be made up of combinations of plane harmonic waves so chosen as to satisfy these boundary conditions. This results in what is known as a *standing-wave pattern*, which depends on the time only through the factor $e^{\pm i\omega t}$ common to all the vectors. Choosing combinations of the second factor $e^{\pm i\mathbf{k}\cdot\mathbf{r}}$ in the form of sines and

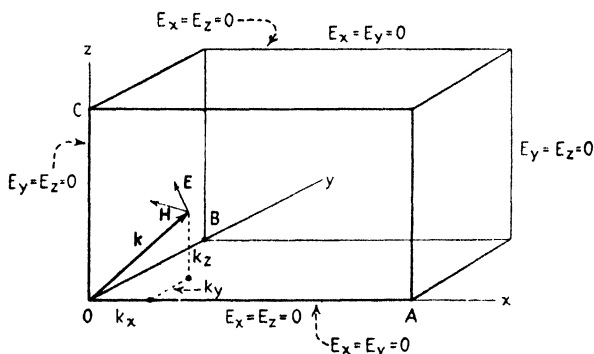



FIG. 16.10.—Boundary conditions and propagation vector in a rectangular cavity.

cosines, it is clear that the boundary conditions will be satisfied by the following components of the electric field:

$$E'_x = E_1 \cos\left(\frac{\pi l x}{A}\right) \sin\left(\frac{\pi m y}{B}\right) \sin\left(\frac{\pi n z}{C}\right)$$
$$E'_y = E_2 \sin\left(\frac{\pi l x}{A}\right) \cos\left(\frac{\pi m y}{B}\right) \sin\left(\frac{\pi n z}{C}\right)$$
$$E'_z = E_3 \sin\left(\frac{\pi l x}{A}\right) \sin\left(\frac{\pi m y}{B}\right) \cos\left(\frac{\pi n z}{C}\right)$$

where l, m, and n are arbitrary integers designating what are called the *modes of oscillation* and the primes refer to the field without the time factor. Thus the components of $\mathbf{k}$ are limited to the values

$$k_x = \frac{l\pi}{A}, \qquad k_y = \frac{m\pi}{B}, \qquad k_z = \frac{n\pi}{C}$$

The only condition on l, m, and n is $k^2 = (\omega/c)^2$, and the only condition on E_1, E_2, and E_3 is that $\mathbf{div}\ \mathbf{E} = 0$ (no charges). These conditions can be written

$$\left(\frac{\omega}{\pi c}\right)^2 = \left(\frac{2}{\lambda_0}\right)^2 = \left(\frac{l}{A}\right)^2 + \left(\frac{m}{B}\right)^2 + \left(\frac{n}{C}\right)^2$$
$$0 = \frac{lE_1}{A} + \frac{mE_2}{B} + \frac{nE_3}{C} \qquad (16.24)$$

It is clear from the equations for the components of $\mathbf{E}'$ that no two of the integers l, m, and n can vanish or all the fields would vanish. Also there are two linearly independent modes of oscillation of the cavity if none of the integers vanish; *i.e.*, a choice of both E_1 and E_2 is necessary to determine E_3. The second equation of Eqs. (16.24) is also the necessary condition that $\mathbf{E}'$ shall be perpendicular to the propagation vector $\mathbf{k}$. The frequency of oscillation can assume only those particular values permitted by the choice of integers for l, m, and n in Eq. (16.24).

The above discussion outlines the nature of the phenomena encountered in a cavity containing radiation. The modes of oscillation are generally designated by subscripts, *i.e.*, E'_{lmn}. If one of the integers vanishes, only one polarization mode is possible; and unless the lengths of the sides of the box are commensurate, only these two polarization modes will have the same frequency and wave length. However, if pairs of sides are commensurate, a plurality of modes may be possible for a particular frequency. Harmonic relationships will exist between certain of the modes, but they will not all be in a harmonic relation to a single fundamental even for a cube. It might also be mentioned that in the case of very high modes a simple approximate expression exists for the number of modes per unit frequency interval. The fractions l/A, m/B, n/C may be considered as coordinates ξ_1, ξ_2, ξ_3; the first equation of Eqs. (16.24) is then that of a sphere of radius

$$\left(\frac{2\nu}{c}\right) = r^2 = \xi_1^2 + \xi_2^2 + \xi_3^2 \qquad \left(\nu = \frac{\omega}{2\pi}\right)$$

The number of unit coordinate cubes corresponding to the lattice of coordinate points determined by $+l$, $+m$, $+n$ lying within the spherical octant of radius r multiplied by 2 (for polarization modes) is then the number of modes of frequency less than ν. This is the volume of the spherical octant, or $8\pi\nu^3/3c^3$. This becomes a very good approximation if λ_0 is less than of the order of a fifth of the dimensions of the box.

In the preceding discussion it has been assumed that the cavity walls are perfectly reflecting. This is a very satisfactory approximation as far as determining the modes of oscillation of the cavity is concerned, but it has been seen in Sec. 16.4 that a small fraction of the incident energy is absorbed by a metal wall. The result of this can be calculated to a good approximation on the assumption that the results of Sec. 16.4 apply to all angles of incidence of a wave on a metal boundary. The total radiant energy in the enclosure can be written in terms of the magnetic field as

$$\int_V (\tfrac{1}{2}\mu_0 H^2 + \tfrac{1}{2}\kappa_0 E^2)\, dv = \int_V \mu_0 H^2\, dv$$

where the integration is over the volume V of the cavity. Then from

Sec. 13.2 and Eq. (16.19′) the Q of the cavity is

$$Q = \frac{2\int_V H^2\,dv}{\delta\int_S H^2\,ds} \tag{16.25}$$

where the integral over the area of the cavity represents the loss to all areas of the walls and H in this integral is the value of the magnetic field in the plane of the surface. As H is a maximum at the surface, it is about twice as great there as its average value throughout the volume. To this crude approximation, which becomes more adequate for high modes of oscillation, the upper integral is about H^2V and the lower is $2H^2S$, where S is the surface area of the cavity; *i.e.*,

$$Q = \frac{V}{\delta S} \tag{16.25′}$$

This simple expression gives the right order of magnitude for Q and shows that it is approximately equal to the ratio of the volume of the cavity to the volume of a surface layer of thickness δ. The quantity Q is thus in general very large. As δ is proportional to the square root of the wave length of the radiation, long wave lengths or low frequencies are attenuated most rapidly per cycle but least rapidly per unit time.[1]

Wave Guides.—Incomplete enclosures in the form of metallic pipes or tubes can be used for the transference of electromagnetic energy. These bear a resemblance both to resonant cavities and to the transmission lines of Sec. 14.5. The region between the conductors of a two-wire transmission line or a coaxial line clearly resembles a cavity in being bounded by metal walls. However, there is a fundamental difference between such lines and a simple pipe. In a simple pipe or enclosure a general Gaussian surface can be contracted to a point without encountering a charge-carrying conductor. This absence of charge within such a region, which is said to be simply connected, results in the second condition of Eqs. (16.24). In the case of the two- conductor line the contraction of a general Gaussian surface which may surround a conductor encounters the charge that it carries, and hence **div E** $= 0$ is not a generally applicable condition. The consequence of this is the type of propagation discussed in Sec. 14.5, in which constant-current transmission is possible and also alternating-current transmission in which the electric and magnetic field vectors are at right angles to the direction of propagation. This essentially means that there is a less stringent limitation on the values of l, m, and n for such lines. All but one of these integers may vanish; the component of k along the line is the only one that must remain finite. Thus to the approximation that the resistivity

[1] General reference: CONDON, *Rev. Mod. Phys.*, **14**, 341 (1942).

of the metal is neglected, both the electric and magnetic fields may be normal to the direction of propagation in a two-conductor line. Such modes of oscillation, which were discussed in Sec. 14.5, are known as *cable modes.*

In the simple pipe containing no internal conductor Eqs. (16.24) continue to apply as for a resonant cavity. However, radiation will be assumed to be introduced at one end and permitted to flow out at the other, so that the boundary conditions of the cavity apply only to the walls of the pipe, not to the ends. Consider a rectangular pipe of dimensions A and B in the x and y directions, respectively, in which an electromagnetic disturbance is propagated in the direction of positive z (Fig. 16.11). It is seen that the equations for the electric field within the

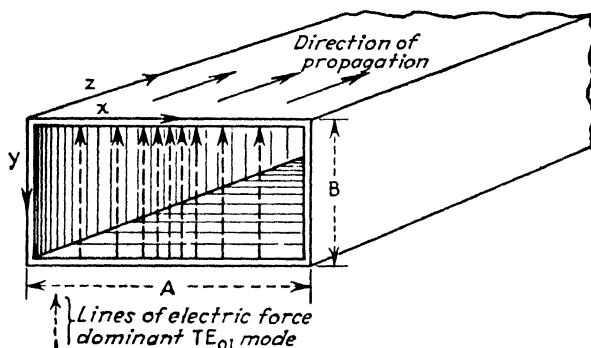



FIG. 16.11.—Propagation of the dominant mode down a rectangular wave guide.

pipe that satisfy the boundary conditions and represent a wave motion in the positive direction are from our earlier discussion

$$E_x = E_1 \cos \frac{\pi l x}{A} \sin \frac{\pi m y}{B} e^{j(\omega t - k_z z)}$$

$$E_y = E_2 \sin \frac{\pi l x}{A} \cos \frac{\pi m y}{B} e^{j(\omega t - k_z z)}$$

$$E_z = E_3 \sin \frac{\pi l x}{A} \sin \frac{\pi m y}{B} e^{j(\omega t - k_z z)} \tag{16.26}$$

The components of the accompanying magnetic field are, of course, given by Eq. 16.4 as

$$\mathbf{H} = \frac{j}{\mu_0 \omega} \operatorname{curl} \mathbf{E}$$

Equations (16.24) remain the same except that k_z is no longer $n\pi/C$; *i.e.*

$$\left(\frac{\omega}{c}\right)^2 = \left(\frac{2\pi}{\lambda_0}\right)^2 = \left(\frac{l\pi}{A}\right)^2 + \left(\frac{m\pi}{B}\right)^2 + k_z^2$$

$$\frac{\pi l}{A} E_1 + \frac{\pi m}{B} E_2 - j k_z E_3 = 0 \tag{16.24'}$$

From the equations for the components of $\mathbf{E}$ it is clear that both l and m cannot vanish or all the fields vanish. It can also be seen on forming $\mathbf{H}_z$ that this component cannot vanish if $\mathbf{E}_3$ does and vice versa. This is in distinction to the cable modes in which both $\mathbf{E}_z$ and $\mathbf{H}_z$ vanish and is the basis for distinguishing two modes of propagation: the TE modes in which there is no component of E along z and the TM modes in which there is no component of H along z. Subscripts are used to distinguish the values of l and m for these modes.

In order that the waves of Eqs. (16.26) shall not be attenuated (wall resistance is neglected), k_z must be real; *i.e.*,

$$\left(\frac{l}{A}\right)^2 + \left(\frac{m}{B}\right)^2 < \left(\frac{2}{\lambda_0}\right)^2$$

The largest value of λ_0 for which this condition is fulfilled is that for which the integer over the smaller dimension vanishes and that over the larger dimension is 1; *i.e.*, $\lambda_{max} = 2A$, $(A > B)$. This is the TE_{01} mode in which $E_x = 0$, and no lower frequency can be propagated freely down the pipe. The longest wave that can be transmitted in the TE_{10} mode, $E_y = 0$, is $\lambda_{max} = 2B$. Thus if λ_0 lies between $2B$ and $2A$, only the TE_{10} mode, known as the *dominant mode*, can be excited. For higher frequencies other propagation modes become possible.

The dominant mode illustrates all of the typical propagation phenomena, and some of its properties will be considered in more detail. On forming $\mathbf{H}$ and setting $m = 0$ and $E_z = 0$

$$\frac{\partial E_y}{\partial z} = j\omega\mu_0 H_x, \qquad \frac{\partial H_x}{\partial z} = \frac{jk_z^2}{\omega\mu_0} E_y \tag{16.27}$$

These are seen to be formally the same as the cable equations (14.26) for current and potential difference with the substitution of E_y for V' and H_x for i'. Thus by analogy γ, the propagation constant, becomes jk_z and the characteristic impedance $\mathbf{Z}_i$ becomes $\mu_0\omega/k_z$. Thus all of the general theory of lines can be taken over in discussing the wave guides. The wave or phase velocity down the guide is from Eq. (16.24′)

$$v' = \frac{\omega}{k_z} = \frac{c}{\left[1 - \left(\frac{l\lambda_0}{2A}\right)^2 + \left(\frac{m\lambda_0}{2B}\right)^2\right]^{1/2}}$$

where λ_0 is the free-space wave length. Writing λ_c the cut-off wave length (λ_{max}) for the TE_{01} mode, the wave velocity for this mode ($m = 0$) becomes

$$v' = \frac{c}{[1 - (\lambda_0/\lambda_c)^2]^{1/2}} \tag{16.28}$$

which is evidently greater than c. In the same way the characteristic impedance for this mode is

$$z_i = \frac{\sqrt{\mu_0/\kappa_0}}{[1 - (\lambda_0/\lambda_c)^2]^{1/2}} \tag{16.29}$$

which is greater than the impedance $\sqrt{\mu_0/\kappa_0}$ for free space. The quantity v', however, is not the velocity with which a signal is propagated down the guide but rather the rate of motion of the interference pattern of traveling waves, which may be thought of as together constituting the TE_{01} mode. Or, in terms of the concepts of optics, the guide represents a dispersive medium, as v' is a function of the frequency. The group velocity, v_g, with which a signal would be transmitted is $d\omega/dk_z$ or $v_g = c[1 - (\lambda_0/\lambda_c)^2]^{1/2}$. This is less than c, and indeed $v_g v' = c^2$. Finally it should be remarked that the finite resistivity of the metal wall leads to some attenuation of the energy transmitted down the guide. This can be calculated from Eq. (16.19′), the energy density within the guide, and the group velocity. For a more complete discussion of these phenomena reference should be made to special treatises.[1]

16.6. Generation of an Electromagnetic Wave.—The preceding sections have been concerned with the propagation of electromagnetic waves under various circumstances. It was seen that in general the existence of partially free charges in the path of the wave results in a transfer of energy from the wave to the charges and the medium with a consequent loss in amplitude of the wave. The inverse phenomenon, namely, the transfer of energy from a changing current or accelerated charge to an electromagnetic wave, will be considered in this section. For this purpose it is necessary to take the complete form of the four fundamental differential equations for free space, including charge and current densities, namely,

$$\text{div } \mathbf{D} = q_v \tag{16.30}$$

$$\text{div } \mathbf{B} = 0 \tag{16.31}$$

$$\textbf{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{16.32}$$

$$\textbf{curl } \mathbf{H} = \mathbf{i}_v + \frac{\partial \mathbf{D}}{\partial t} \tag{16.33}$$

It is more convenient to deal with the scalar and vector potentials than with the electric and magnetic vectors of the wave. However, these potentials have been defined only for static charges or steady currents.

[1] SOUTHWORTH, *Bell System Tech. J.*, **15**, 284 (1936); CARSON, MEAD, and SCHELKUNOFF, *Bell System Tech. J.*, **15**, 310 (1936); SLATER, "Microwave Transmission," McGraw-Hill Book Company, Inc., New York, 1942; MONTGOMERY, DICKE, and PURCELL, "Principles of Microwave Circuits," McGraw-Hill Book Company, Inc., New York, 1948.

Hence it is first necessary to extend their definitions to the case in which they are functions of the time. Substituting $\mathbf{B} = \mathbf{curl\,A}$ in Eq. (16.32), it is found that

$$\mathbf{curl}\left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right) = 0$$

Since any vector function for which the **curl** vanishes is the **grad** of a suitable potential, it is evident that the quantity in parenthesis can be written as the gradient of a scalar quantity. To agree with electrostatics the parenthesis must be equal to minus the gradient of the scalar potential, or

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \mathbf{grad}\ V \tag{16.34}$$

Substituting this expression in Eqs. (16.30) and (16.33) and using the conditions for free space, $\mathbf{D} = \kappa_0\mathbf{E}$ and $\mathbf{B} = \mu_0\mathbf{H}$, the following equations are obtained:

$$-\nabla^2 V - \operatorname{div}\frac{\partial \mathbf{A}}{\partial t} = \frac{q_v}{\kappa_0}$$

$$-\nabla^2\mathbf{A} + \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} + \mathbf{grad}\left(\operatorname{div}\mathbf{A} + \frac{1}{c^2}\frac{\partial V}{\partial t}\right) = \mu_0\mathbf{i}_v$$

where $\mathbf{grad}\ \operatorname{div}\mathbf{A} - \nabla^2\mathbf{A}$ has been written for $\mathbf{curl\ curl\ A}$ and $1/c^2$ for $\kappa_0\mu_0$. As only the **curl** of **A** and its partial derivative with respect to t have been defined, its divergence is still unspecified. For simplicity, this is generally defined in such a way as to make the parenthesis in the second equation vanish, or

$$\operatorname{div}\mathbf{A} + \frac{1}{c^2}\frac{\partial V}{\partial t} = 0 \tag{16.35}$$

With this definition the two equations become:

$$-\nabla^2 V + \frac{1}{c^2}\frac{\partial^2 V}{\partial t^2} = \frac{q_v}{\kappa_0} \tag{16.36}$$

$$-\nabla^2\mathbf{A} + \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = \mu_0\mathbf{i}_v \tag{16.37}$$

These differ from the fundamental equations of electrostatics and magnetostatics only in the second term on the left-hand side. But it has previously been seen that this is just the term that is responsible for the wave nature of this type of equation. The general solutions reflect these relationships and may be written as follows:[1]

[1] LORENTZ, H. A., "The Theory of Electrons," B. G. Teubner, Leipzig, 1908, Appendix IV; MASON and WEAVER, "The Electromagnetic Field," p. 282, University of Chicago Press, 1929; STRATTON, "Electromagnetic Theory," McGraw-Hill Book Company, Inc., New York, 1941.

$$V = \frac{1}{4\pi\kappa_0}\int \frac{(q_v)}{r}\,dv$$

$$\mathbf{A} = \frac{\mu_0}{4\pi}\int \frac{(\mathbf{i}_v)}{r}\,dv$$

It will be noted that these resemble the fundamental solutions of the static electric and electromagnetic differential equations encountered in earlier chapters, [Eqs. (1.10) and (9.13)]. The parentheses around the charge and current densities indicate that in performing the integration to obtain the potentials at a time t the charges and currents prevalent at an earlier time $\left(t - \frac{r}{c}\right)$ must be taken. This is reasonable since it has already been seen that an electromagnetic disturbance is propagated with a velocity c and a time r/c must elapse before the effect of any alteration in q_v or $\mathbf{i}_v$ is observed a distance r away. These expressions for V and $\mathbf{A}$ are known as *retarded potentials.* The general solution will not be considered further, but our attention will be confined to a few special cases.

Consider a current i flowing in an infinitesimal length of wire represented by the vector $\mathbf{dl}$. Near the wire the vector potential must reduce to the form of the Biot-Savart law which is seen to be the case on substituting $i\,\mathbf{dl}$ for $\mathbf{i}_v\,dv$ in the expression for $\mathbf{A}$

$$\mathbf{dA} = \frac{\mu_0}{4\pi}\frac{i}{r}\,\mathbf{dl}$$

At a distant point

$$\mathbf{dA} = \frac{\mu_0}{4\pi}\frac{i[t - (r/c)]}{r}\,\mathbf{dl}$$

where $i[t - (r/c)]$ is written for (i). The magnetic induction is given by **curl A** or, since **dl** is not a function of the coordinates

$$\begin{aligned}\mathbf{dB} &= \mathbf{curl\ dA} \\ &= \frac{\mu_0}{4\pi}\,\mathbf{grad}\left[\frac{i[t - (r/c)]}{r}\right] \times \mathbf{dl} \\ &= \frac{\mu_0}{4\pi}\frac{\partial}{\partial r}\left[\frac{i[t - (r/c)]}{r}\right]\mathbf{r}_1 \times \mathbf{dl} \\ &= \frac{\mu_0}{4\pi}\left[\frac{i[t - (r/c)]}{r^2} + \frac{1}{rc}\frac{\partial i(t - r/c)}{\partial t}\right]\mathbf{dl} \times \mathbf{r}_1\end{aligned}$$

The first term is the ordinary Biot-Savart law for the magnetic induction of a steady current. It falls off inversely as the square of the distance from the wire. The second term, which does not vanish if the current is changing, decreases only as the inverse first power of r and hence becomes predominant at great distances. This is known as the radiation term

of the induction. It is proportional to the rate of change of current or to the acceleration of a charge. The electric field can be found from Eqs. (16.34) and (16.35), but its limiting values very close to the wire or at great distances can be found more simply. Since **E** is continuous in the wire and in the space outside, the neighboring value of the field strength is equal to the electric field in the wire. The magnitude of **E** at a great distance, or the radiation field, is given in terms of the induction by Eq. (16.7) as cB. It is perpendicular to the induction and to the direction of propagation $\mathbf{r}_1$, *i.e.*, $\mathbf{E} = c\mathbf{B} \times \mathbf{r}_1$. Assuming that (i) is a simple periodic function, $(i) = i_0 e^{j\omega[t-(r/c)]}$, the expression for $\mathbf{dB}$ becomes

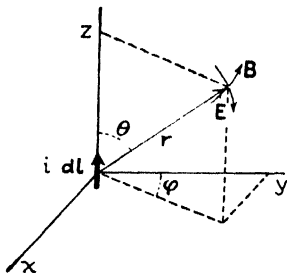



FIG. 16.12.—Electric and magnetic radiation vectors from a current element i dl.

$$\mathbf{dB} = \frac{\mu_0}{4\pi}\left(\frac{1}{r^2} + \frac{j\omega}{rc}\right)(i)\ \mathbf{dl} \times \mathbf{r}_1$$

The second term is much greater in absolute magnitude than the first if $r >> \lambda$; thus the Biot-Savart term can in general be neglected at distances greater than a few hundred wave lengths. Considering that all observations will be made at greater distances than this, the radiation terms can be written for the induction and field strength as

$$\mathbf{dB} = \frac{\mu_0}{2r\lambda}(i)\ \mathbf{dl} \times \mathbf{r}_1 \tag{16.38}$$

$$\mathbf{dE} = \frac{c\mu_0}{2r\lambda}(i)\ (\mathbf{dl} \times \mathbf{r}_1) \times \mathbf{r}_1 \tag{16.39}$$

Since $\mathbf{dB}$ is perpendicular to $\mathbf{r}_1$ and $\mathbf{dl}$ and proportional to the sine of the angle between them, the lines of induction can be thought of as the parallels of latitude on a sphere, with the exception that the lines are closely spaced at the equator but less dense near the polar axis which is in the direction of **dl**. The lines of electric force, being perpendicular to $\mathbf{r}_1$ and the induction with a density proportional to the latter, are in the direction of circles of longitude through the polar axis. The relations between these vectors are indicated in Fig. 16.12.

An isolated current element i **dl** can be thought of as an electric dipole of constant length **dl** with terminal charges $\pm q = \pm q_0 e^{j\omega t}$, writing $q\ \mathbf{dl} = \mathbf{p}$, $i\ \mathbf{dl} = d\mathbf{p}/dt = j\omega\mathbf{p}$. Hence this type of radiation in terms of either i **dl** or **p** is called dipole radiation. As the wave spreads out radially from the dipole, it carries energy with it. This rate of loss of energy can be determined by means of Poynting's vector, Eq. (16.8). Writing $i\ \mathbf{dl} = -\omega\mathbf{p}_0 \sin \omega t$, and θ for the polar angle between **p** and $\mathbf{r}_1$

$$\mathbf{N} = \mathbf{E} \times \mathbf{H}$$

or

$$\mathbf{dN} = \frac{1}{\mu_0}\mathbf{dE} \times \mathbf{dB}$$

$$dN = c\mu_0\left(\frac{p_0\omega}{2r\lambda}\right)^2 \sin^2\theta \sin^2\omega t$$

Since this is proportional to the square of the sine of the polar angle, it is evident that most of the energy is in the region of the equatorial plane and that none is emitted in the direction of the dipole itself. The average rate of flow of energy per unit area is obtained by taking the average of $\sin^2 \omega t$ over a complete period, which is $\frac{1}{2}$. To get the average rate of loss of energy in all directions the expression must be integrated over a sphere of radius r, or

$$dP = \tfrac{1}{2}\int_0^\pi 2\pi\, dNr^2 \sin\theta\, d\theta$$

On making the substitution $x = \cos\theta$ this expression becomes

$$dP = \pi c\mu_0\left(\frac{p_0\omega}{2\lambda}\right)^2 \int_{-1}^{1}(1 - x^2)dx$$

$$= \frac{\pi c\mu_0}{3\lambda^2}(p_0\omega)^2 = \frac{\mu_0}{12\pi c}p_0^2\omega^4 \qquad (16.40)$$

Thus the power radiated by a dipole is proportional to the square of the amplitude and to the fourth power of the frequency.

The most intense light emitted by atoms and molecules is of the dipole type. The energy radiated corresponds to a loss of energy by the atomic system and an eventual cessation of the type of motion responsible for the radiation. A more detailed discussion of this problem would involve the introduction of the quantum theory which must be used in dealing with these microscopic systems, and this is beyond the scope of our treatment. To discuss the radiation from a current in a wire, it is more convenient to replace $\mathbf{p}_0\omega$ of Eq. (16.40) by $i_0\,\mathbf{dl}$. Inserting the numerical values of the constants and writing y for $\mathbf{dl}/\lambda$ the equation becomes

$$dP = 395y^2i_0^2$$
$$= 790y^2i_e^2$$

where i_0 is the maximum and i_e the rms. or effective current in amperes and dP is the power in watts. On comparing this with the ordinary ohmic power loss, it is seen that the effective resistance exhibited by the circuit is $790y^2$ ohms. This is known as the *radiation resistance*. The previous discussion of high-frequency currents in wires has shown that in general the current is not the same at all points. However, the approximation involved in considering it uniform is valid for small

values of $\mathbf{dl}/\lambda$ or y and the expression may be used to calculate the radiation contribution to the resistance of a wire for values of y from 0 to 0.1.

The electromotive force induced in a passive wire by an incident electromagnetic wave can be readily calculated from the condition that the electric field is continuous across the boundary of the wire. Con-

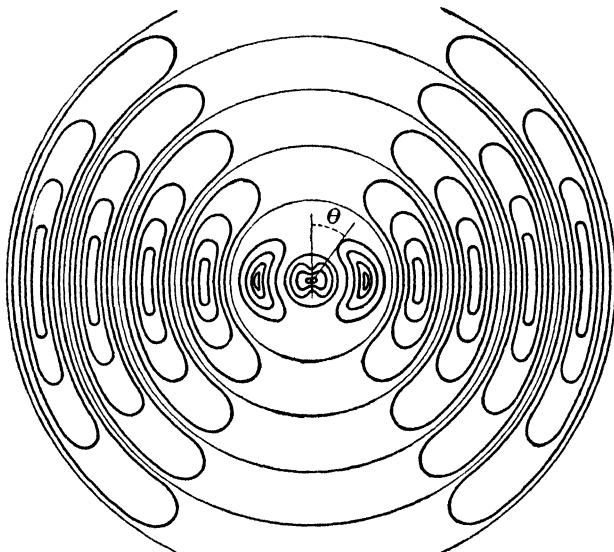



FIG. 16.13.—Schematic instantaneous representation of the lines of electric force in a plane containing a radiating dipole.

sider a radiating wire element of length $\mathbf{dl}_1$ which is carrying a current i_1; the electric field at a distance $\mathbf{r}$ is given by Eq. (16.39) as

$$dE = \frac{c\mu_0}{2\lambda r^3}(i_1)(\mathbf{dl}_1 \times \mathbf{r}) \times \mathbf{r}$$

If a wire element of length $\mathbf{dl}_2$ is situated at $\mathbf{r}$, the emf. generated in it is $\mathbf{dE} \cdot \mathbf{dl}_2$; calling this $\mathcal{E}_2$, it is seen that

$$\mathcal{E}_2 = \frac{c\mu_0}{2\lambda r^3}(i_1)(\mathbf{dl}_1 \times \mathbf{r}) \times \mathbf{r} \cdot \mathbf{dl}_2$$

$$= -\frac{c\mu_0}{2\lambda r^3}(i_1)(\mathbf{r} \times \mathbf{dl}_1) \cdot (\mathbf{r} \times \mathbf{dl}_2)$$

This expression is entirely symmetrical in $\mathbf{dl}_1$ and $\mathbf{dl}_2$, so if the element $\mathbf{dl}_2$ had been carrying the current i_1, an identical emf. would have been generated in the element $\mathbf{dl}_1$. This reciprocity theorem is an extremely useful one and simplifies many calculations. In deriving it the tacit assumption has been made that there are no unsymmetrical effects that

would depend on the direction of the vector **r** such as changes in the nature of the polarization of the wave. This assumption would be violated by transmission through the ionosphere, for instance, and the reciprocity theorem cannot be applied under those conditions.[1]

One of the simplest and most useful circuits for the detection of radio waves is the loop antenna. In its simplest form this is merely a closed loop of wire, but in general the loop is composed of a number of turns, electrostatically shielded from undesired disturbances, in series with a parallel resonant circuit, across which the greater part of the induced emf. appears. Such a loop is indicated schematically in Fig. 16.15. If **E** is the electric field strength associated with the wave and **dl** is an element of length of the loop, the induced emf. is given by

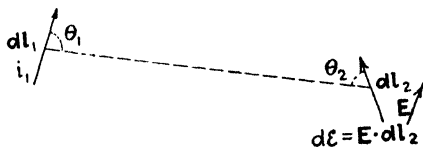



FIG. 16.14.—Illustration of the reciprocal relation existing between the inducing current and the induced emf. for two infinitesimal segments.

$$\mathcal{E} = \oint \mathbf{E} \cdot \mathbf{dl} = \int (\text{curl } \mathbf{E}) \cdot \mathbf{ds} = -\int \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds}$$

where **B** is the magnetic induction and **ds** is an element of area of the loop. If the linear dimensions of the loop are small in comparison

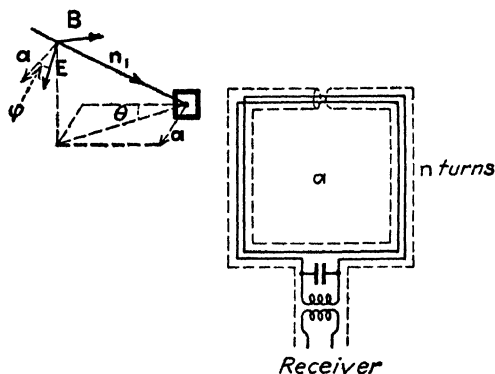



FIG. 16.15.—Loop antenna.

with the length of the wave, **B** is approximately uniform over the area of the loop. Writing **a** for the vector representing the area of the loop and differentiating **B**, the induced emf. becomes

$$\mathcal{E} = -j\omega \mathbf{B} \cdot \mathbf{a}$$

[1] For a discussion of reciprocal theorems see Carson, *Proc. I.R.E.*, **17**, 952 (1929).

Or in terms of the electric field

$$\begin{aligned} \mathcal{E} &= -\frac{j\omega}{c}\mathbf{n} \times \mathbf{E} \cdot \mathbf{a} \\ &= \frac{2\pi a}{\lambda} \cos\theta \sin\phi\, E \end{aligned} \tag{16.41}$$

where the phase difference indicated by j is neglected and the angles are those indicated in Fig. 16.15. It is evident from any of these forms that the induced emf. is greatest when the magnetic induction of the wave is in the direction of **a**, *i.e.*, normal to the plane of the loop. Thus the plane of the loop for a maximum signal coincides with the plane determined by the wave normal and the electric vector. The direction of the wave normal is not determined, but if, for instance, it is known that the wave is coming directly from some point on the surface of the earth, the orientation of the loop determines the direction of the transmitting station. By utilizing this directional property, aircraft and ships equipped with loop antennas can determine their position by taking bearings on radio beacons that are known to be located at certain points.

The efficiency of a loop as a receiver or radiator in comparison with a straight wire can be obtained from Eq. (16.41). The emf. induced in a wire of length $\mathbf{l}$ is $\mathbf{E} \cdot \mathbf{l}$, and since $\mathbf{n}$ is a unit vector perpendicular to $\mathbf{E}$, Eq. (16.41) can be written $(2\pi/\lambda)\mathbf{E}' \cdot \mathbf{a}$, where $\mathbf{E}'$ is a vector of length E but rotated $\pi/2$ about the wave normal, *i.e.*, in the direction of **B**. The ratio of the emf. induced in a loop to that induced in a wire is then

$$\frac{\mathcal{E}_l}{\mathcal{E}_w} = \frac{2\pi a}{\lambda l}$$

for the optimum or corresponding orientations. If the linear dimensions of the loop are of the order of the length of the wire, this ratio is of the order of l/λ. Since it has been assumed that the loop dimensions are small compared with the wave length, this ratio is small and the loop is evidently a poor quantitative receiver. By the reciprocity theorem, the emfs. induced in two elements for a certain field strength are proportional to the field strengths that would be produced at a great distance for equal currents flowing in the two elements. Hence the ratio of the radiation field produced by a loop to that produced by a straight wire is also equal to $2\pi a/\lambda l$; the relative positions of the electric and magnetic vectors are, of course, rotated through an angle $\pi/2$ about the wave normal. The power radiated by a loop is less by the square of this ratio than the power radiated by a wire. As an example consider a wire 1 m. long and a circular loop 1 m. in diameter, each carrying a current which oscillates with a frequency of a million cycles per second. The radiation resistance of the wire is $790/(300)^2 = 0.88 \times 10^{-2}$ ohm. The resistance of the loop is less by the factor 2.5×10^{-4} or about

2.2×10^{-6} ohm. Hence it is obvious that the radiation contribution to the resistance of ordinary coils can be neglected even at very high frequencies.

16.7. Radiation from Systems in Which the Current Is Not Uniform. At very high frequencies the wave length of the radiation becomes of the order of the length of the wire acting as radiator, and it is no longer legitimate to consider that the current at any instant is uniform throughout the antenna. A complete discussion of practical antenna systems and their characteristics is beyond the scope of this treatment and will be found in books on radio engineering. The presence of the surface of the earth has a very important effect on radiating systems. This can be taken into account to a first approximation by assuming that the earth's surface is a perfectly conducting plane. On this assumption the simple electrostatic theory of image charges can be used, and an antenna above the surface of the earth can be considered as being accompanied by its mirror image below the surface. The radiation in the free space above the earth is that which would be due to the two antennas in free space. The validity of the approximation can be improved by placing a highly conducting area or counterpoise over the surface of the ground in the neighborhood of the base of the antenna. If the approximation is valid, the problem is reduced to that of calculating the radiation fields due to the two antennas. The discussion can be conveniently divided into two parts, that dealing with those antennas in which owing to the terminal conditions standing waves exist, and that concerned with antennas terminated by their characteristic impedance so that the waves are not reflected but absorbed at the ends.

In the first class is the half wave or doublet antenna. This is a straight wire which to an adequate approximation is exactly $\lambda/2$ long. When excited by the proper frequency, the ends are potential antinodes and the center is a current antinode. It is analogous to an organ pipe open at both ends, oscillating in its fundamental mode. The solution of the problem will also include that of a quarter-wave antenna erected vertically above a perfectly conducting earth, since the lower half of a half-wave antenna is the mirror image of the upper half. Neglecting the vectorial nature of Eq. (16.39), which is legitimate since at great distances the forces from the different elements of the antenna are colinear and can be added algebraically, the field strength at a distance r due to an element dl carrying a current $i = i' \sin \omega\left(t - \frac{r}{c}\right)$ becomes on putting in the numerical factors

$$dE = 188.5 \sin \theta \frac{i' \sin \omega\left(t - \frac{r}{c}\right)}{\lambda r} dl$$

The assumed current distribution in the half-wave antenna implies that i' can be written $i' = i_0 \cos 2\pi l/\lambda$, where l is measured from the center of the wire. The appropriate value for r depends on the direction of observation and the position of the element dl along the antenna. From Fig. 16.16 $r = r_0 - l \cos \theta$, where r_0 is the distance from the center of the antenna to the distant point of observation. The difference between r and r_0 is small for all points on the antenna and r may be set equal to r_0 in the denominator, but the difference must be taken into account in the periodic part of the function. On grouping together elements situated symmetrically $\pm l$ on either side of the center of the antenna, the electric vector due to the half-wave antenna becomes

FIG. 16.16.—Analysis of the radiation from a half-wave antenna.

$$E = 188.5 i_0 \frac{\sin\theta}{\lambda r_0} \int_0^{\frac{\lambda}{4}} \left\{ \sin \omega\left(t - \frac{r_0}{c} + \frac{l \cos\theta}{c}\right) + \sin \omega\left(t - \frac{r_0}{c} - \frac{l\cos\theta}{c}\right)\right\} \cos \frac{2\pi l}{\lambda}\, dl$$

$$= 377 i_0 \frac{\sin \omega\left(t - \frac{r_0}{c}\right)}{\lambda r_0} \sin\theta \int_0^{\frac{\lambda}{4}} \cos\frac{2\pi l}{\lambda} \cos \frac{2\pi l \cos\theta}{\lambda}\, dl$$

On writing the integrand in terms of the sum and difference between the two angles, the integration can be performed immediately, leading to

$$E = 60 \frac{i_0}{r_0} \frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta} \sin\omega\left(t - \frac{r_0}{c}\right) \tag{16.42}$$

E is the field strength in volts per meter and i_0 is the current in amperes at the center of the antenna; these may, of course, be either the maximum or effective values of the quantities. A polar diagram of the field strength given by this equation is shown in Fig. 16.17. The length of the vector from the origin to any point on the curve is proportional to the magnitude of the field strength in that direction from the antenna. The direction of the field in space is, of course, given by the vector equation (16.39). Since E is independent of ϕ, the pattern is the same in all azimuths and Fig. 16.17 can be considered as a section through a torus. The analogous pattern for a wire carrying a uniform current would resemble this pattern, but the curves would be circles tangent to one another at the origin.

The power radiated by a half-wave antenna, and hence its radiation resistance, can be calculated by means of Poynting's vector. From

Eq. (16.8) the power radiated per unit area is $2.65 \times 10^{-3} E^2$; inserting the value of E given by Eq. (16.42) and taking the time average of the periodic term, which introduces a factor of one-half, the power radiated through a sphere of radius r_0 becomes

$$P = 4.77\frac{i_0^2}{r_0^2}\int_0^\pi \frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta}2\pi r_0^2 \sin\theta\, d\theta$$

The value of the integral between these limits is $2.438\pi r_0^2$ which gives the total average power as

$$P = 36.57 i_0^2 = 73.14 i_e^2 \tag{16.43}$$

Thus the radiation resistance of a half-wave antenna is 73.14 ohms. This does not mean that the impedance as observed by a source of emf.

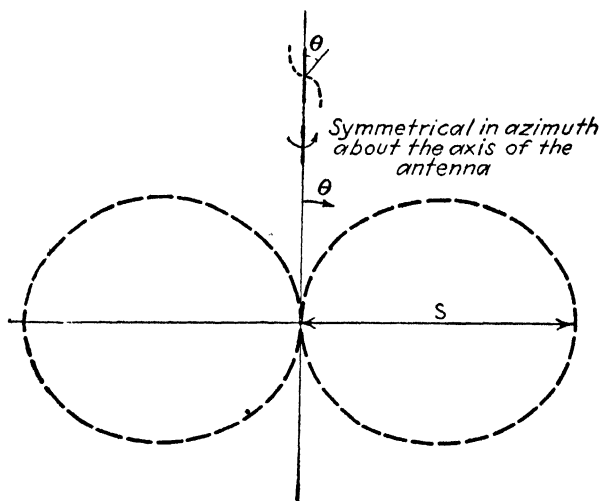



FIG. 16.17.—Radiation pattern of a half-wave antenna.

at the center of the antenna is purely resistive, but this is merely the resistive component of the impedance. There is, as a matter of fact, a reactive component, which may be calculated on certain assumptions to be 42.5 ohms; this, of course, does not affect the power relations. For a more complete account of the circuit relations in radiating systems reference should be made to the literature.[1]

[1] CARTER, *Proc. I.R.E.*, **20**, 1004 (1932); SCHELKUNOFF, "Electromagnetic Waves," D. Van Nostrand Company, Inc., New York, 1943; KING, MIMNO, and WING, "Transmission Lines, Antennas, and Wave Guides," McGraw-Hill Book Company, Inc., New York, 1945; TERMAN, "Radio Engineering," McGraw-Hill Book Company, Inc., New York, 1947.

The half-wave antenna is directive in the sense that most of the power is radiated in the region of the equatorial plane, which is also true of a wire carrying a uniform current. Much more pronounced directional effects can be produced by a group of half-wave antennas suitably spaced and excited in the proper phase. Consider a group of such antennas spaced within a few wave lengths of one another, oriented in the same direction and carrying currents of equal magnitude. The coefficient of the periodic term of Eq. (16.42) is then common to them all. However, the currents will not be assumed to be in phase with one another and the spacial distribution of the antennas will induce an additional apparent phase shift in the direction of observation. On abbreviating $60\frac{i_0}{r_0}\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}$ to S' and writing the periodic term in the exponential form, Eq. (16.32) becomes

$$E = S'e^{j\omega\left(t+T_i-\frac{r_0-\mathbf{R}_i\cdot\mathbf{r}_1}{c}\right)}$$

where $\mathbf{R}_i$ is the vector representing the position of the ith antenna, T_i is its temporal phase, and $\mathbf{r}_1$ is a unit vector in the direction of observation. The total field strength due to n antennas would then be written

$$E_n = S'e^{j\omega\left(t-\frac{r_0}{c}\right)}\sum_{i=1}^{i=n} e^{j\omega\left(T_i+\frac{\mathbf{R}_i\cdot\mathbf{r}_1}{c}\right)}$$

Limiting the discussion to the case in which the antennas are uniformly spaced a distance d apart along a line with a relative phase lag $\varphi = \omega T$ between neighboring ones, the expression becomes

$$E_n = S'e^{j\omega\left(t-\frac{r_0}{c}\right)}\sum_{m=1}^{m=n} e^{j\left(\varphi+\frac{(2\pi d\cos\psi)}{\lambda}\right)}$$

where ψ is the angle between the direction of d and the direction of observation, as shown in Fig. 16.18.

This summation can be performed algebraically or graphically; Fig. 16.18 illustrates the latter method. Writing α for the quantity $\left[\varphi + \frac{(2\pi d\cos\psi)}{\lambda}\right]$, it is evident that the sum of n equal vectors of length F making successive angles α with one another is given by

$$E_n = F\frac{\sin(n\alpha/2)}{\sin(\alpha/2)} \tag{16.44}$$

where F is written for $S'e^{j\omega\left(t-\frac{r_0}{c}\right)}$ and α is a function of φ, ψ, d, and λ.

The maxima of the radiation pattern are found by equating the $dE_n/d\alpha$ to zero, this yields the equation

$$n \tan \frac{\alpha}{2} = \tan \frac{n\alpha}{2}$$

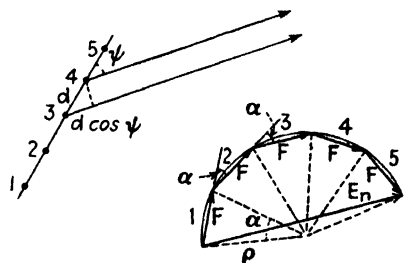
which can be solved approximately by the use of a table of tangents. If φ is known the angle ψ corresponding to the maxima can be determined. It is here assumed that ψ lies in the equatorial plane of the half-wave antenna; otherwise ψ and θ are not independent. The zeros of the radiation pattern occur at $\sin (n\alpha/2) = 0$, except for the case in which $\sin \frac{\alpha}{2}$ is also zero which corresponds to the principal maximum. For $\alpha = 0$, the ratio of the sines is n, hence the principal maximum value of E_n is nF, which occurs at the angle determined by $-\varphi = (2\pi d \cos \psi)/\lambda$. The neighboring zeros on either side are those for which $n\alpha/2 = \pm\pi$. Thus the angular width of the principal maximum is determined by this condition through the dependence of α on ψ. If all the radiators are in phase φ is zero and the principal maximum occurs at $\psi = \pi/2$. If n is large, the width of the principal maximum is small, for if $\delta\alpha$ is the difference between the two values of α corresponding to the neighboring minima

$F = 2\rho \sin \alpha/2 \quad E_n = 2\rho \sin n\alpha/2$
or $\quad E_n = F \dfrac{\sin n\alpha/2}{\sin \alpha/2}$

FIG. 16.18.—Graphical analysis of the radiation pattern from a group of antennas.

$$\delta\alpha = \frac{4\pi}{n}$$

or in terms of ψ (for $\varphi = 0$)

$$\delta\psi = \frac{2\lambda}{nd}$$

Since the total length of the array is $(n - 1)d$, if n is large the angular width of the pattern is approximately twice the ratio of the wave length to the width of the antenna array.

The radiation patterns for a few simple specific cases are shown in the accompanying figures. Figures 16.19 and 16.20 illustrate the radiation patterns from two and five half-wave antennas normal to their common equatorial plane and a distance $\lambda/2$ apart. These figures indicate the way in which the sharpness of the principal maximum increases with n. The patterns are seen to be symmetrical about the plane containing the

$$E_n = 2S\frac{\sin(\pi \cos\psi)}{\sin(\pi/2 \cos\psi)}$$

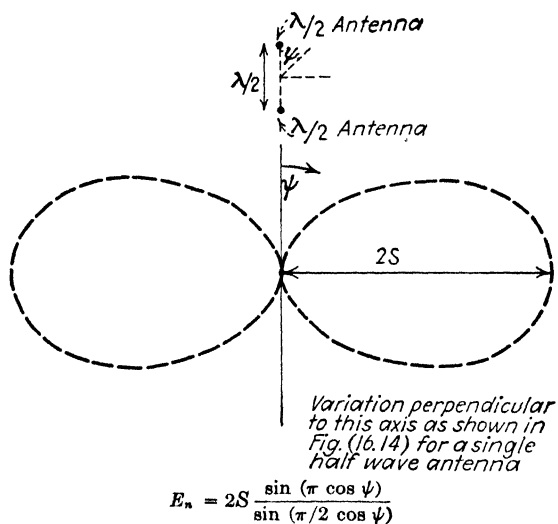



FIG. 16.19.—Radiation pattern of two half-wave antennas in phase and one half wave apart in space.

$$E_n = 5S\frac{\sin(5\pi/2 \cos\psi)}{\sin(\pi/2 \cos\psi)}$$

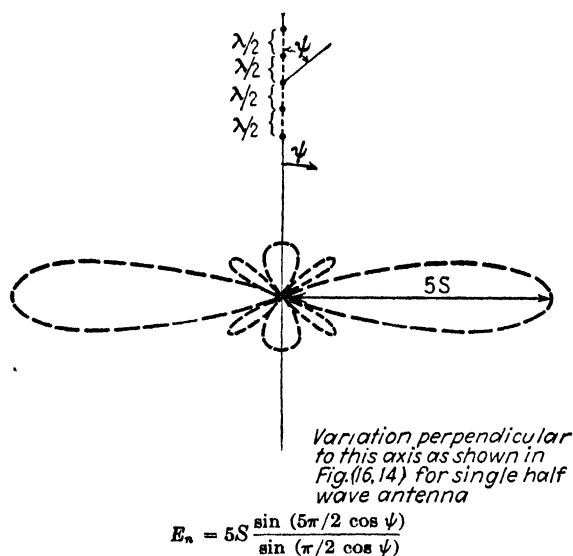



FIG. 16.20.—Radiation pattern of five half-wave antennas in phase and spaced one-half wave length apart.

antennas. Figure 16.21 illustrates the unsymmetrical pattern due to two antennas $\lambda/4$ apart in space and differing in phase of excitation by $\pi/2$. In this case α is equal to $\pi/2$ $(\cos\psi - 1)$ and Eq. (16.44) becomes

$$E_n = 2F \cos\frac{\pi}{4}(1 - \cos\psi)$$

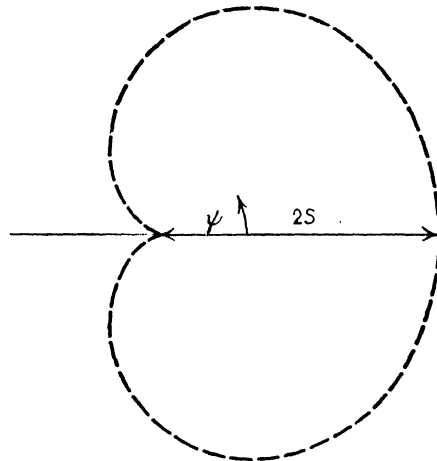
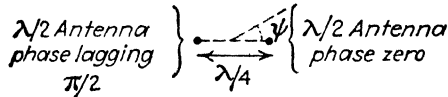
In this case, as can be seen from the figure, the majority of the radiation is in the direction of a vector from the lagging to the leading antenna.

$$E_n = 2S \cos\frac{\pi}{4}(1 - \cos\psi)$$

FIG. 16.21.—Radiation pattern of two half-wave antennas one-quarter wave length apart and $\pi/2$ out of phase.

If each of the antennas of Figs. 16.19 or 16.20 were backed at a distance $\lambda/4$ by an antenna lagging in phase by $\pi/2$, the effect would evidently be to multiply the radii of those figures by the radii of Fig. 16.21, suppressing almost entirely the radiation in the direction of the lagging antenna and multiplying the amplitude of the principal forward maximum by 2. This may be accomplished approximately simply by putting unexcited dummy antennas at the proper distance behind the primary ones for the currents induced in these will be in the proper phase to suppress the back wave. Any number of standing waves on a wire can also be considered as a symmetrical arrangement of colinear half-wave antennas. In this case the phase difference is π and the spacing is $\lambda/2$. Also the angle ψ is the same as θ, and combining Eqs. (16.42) and (16.44).

$$E_n = 60\frac{i_0}{r_0}\sin\cdot\omega\left(t - \frac{r_0}{c}\right)\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)\sin\frac{n\pi}{2}(1+\cos\theta)}{\sin\theta\sin\frac{\pi}{2}(1+\cos\theta)}$$

which reduces to

$$E_n = S\frac{\sin(n\pi/2\cos\theta)}{\sin\theta} \qquad n \text{ even}$$

or

$$E_n = S\frac{\cos(n\pi/2\cos\theta)}{\sin\theta} \qquad n \text{ odd}$$

where S is written for $60\frac{i_0}{r_0}\sin\omega\left(t - \frac{r_0}{c}\right)$. Two representative patterns for two and five half waves are shown in Figs. 16.22 and 16.23, respectively. The patterns are symmetrical in azimuth about the axis of the wire; hence the complete spacial polar surface would be formed by rotating the figure about this axis. For a more complete discussion of these directive arrays reference should be made to the technical literature.[1]

$$E_n = S\frac{\sin(\pi\cos\psi)}{\sin\psi}$$

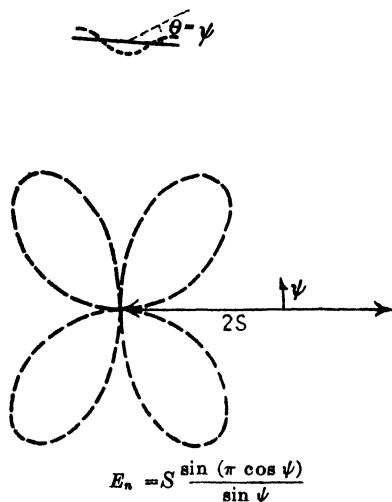



FIG. 16.22.—Radiation pattern due to a full-wave antenna or two colinear half-wave antennas π apart in phase and half a wave length apart in space.

The second important group of high-frequency antennas are those which are terminated by their proper characteristic impedance so that disturbances are not reflected from the terminations and hence no standing waves exist. One or two illustrative antennas of this type will be discussed under certain simplifying conditions; for a more detailed account reference should be made to the current literature.[2] Consider a plane wave incident on a wire of length l with its electric vector in the plane determined by the wave normal and l. This is illustrated in Fig. 16.24. The emf. generated in a length dx of the wire at a point x is $E_x\,dx$ $\cos\phi$, where E_x is the field strength at the point x. Since l is not small in comparison with the wave length, it is necessary to take into account the

[1] WILMOTTE and MCPETRIE, *J.I.E.E.*, **66**, 949 (1928); CARTER, HANSELL, and LINDENBLAD, *Proc. I.R.E.*, **19**, 1773 (1931).

[2] BRUCE, *Bell System Tech. J.*, **10**, 656 (1931); BRUCE, BECK, and LOWRY, *Bell System Tech. J.*, **14**, 135 (1935); FOSTER, *Proc. I.R.E.*, **25**, 1327 (1937).

variation in the phase of E in calculating E_x. Writing the periodic term in its exponential form, it is seen that the phase factor a distance x from R is given by $e^{(j2\pi/\lambda)x \sin \phi}$. There is an additional phase lag before this impulse reaches R along the wire given by $e^{-(j2\pi/\lambda)x}$. It is assumed that there is no reflection from the upper end of the wire; hence the integral of the product of these terms along the wire gives the total effective emf.

$$E_n = S \frac{\cos(5\pi/2 \cos \psi)}{\sin \psi}$$

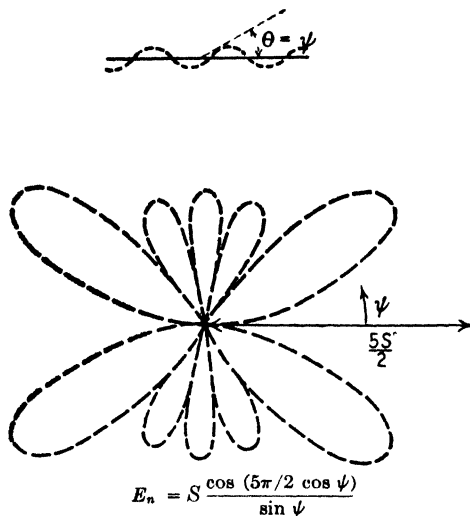



FIG. 16.23.—Radiation pattern due to a five-half-waves antenna or to five colinear half-wave antennas. π apart in phase and half a wave length apart in space.

Calling the total effective impedance of the system $\mathbf{z}$, the current through R is given by

$$\begin{aligned} \mathbf{i}_R &= \frac{E}{\mathbf{z}} \int_0^l \cos \phi e^{j\frac{2\pi}{\lambda}(\sin \phi - 1)x} \, dx \\ &= -\frac{jE\lambda \cos \phi}{2\pi \mathbf{z}(1 - \sin \phi)} \left(1 - e^{j\frac{2\pi}{\lambda}(\sin \phi - 1)l}\right) \end{aligned} \tag{16.45}$$

This current is principally sensitive to the exponential term and will have its greatest value when $e^{j\frac{2\pi}{\lambda}(\sin\phi - 1)l} = -1$ which corresponds to $(l/\lambda)(\sin \phi - 1) = m/2$, where m is odd. This means that the current through R is a maximum when the difference between the number of waves along the ground under the wire and the number along the wire is an odd multiple of $\frac{1}{2}$. Hence the tilted wire has directive properties in the sense that the current through R varies with the angle of approach of the wave. A more useful antenna can be produced by adding a second tilted wire properly terminated to suppress reflection at its lower end, as shown in Fig. 16.25. The analysis is exactly the same for wire b, and for wire a the sign of $\cos \phi$ is evidently negative and the integration

is from l to $2l$. Performing the integration and recalling that the impedance is twice that for a single wire, the current through R reduces to

$$i_R = \frac{jE\lambda \cos \phi}{2z2\pi(1 - \sin \phi)}[1 - e^{j\frac{2\pi}{\lambda}(\sin\phi - 1)l}]^2 \tag{16.46}$$

The principal difference between Eqs. (16.45) and (16.46) is that in the latter the second term is squared. Thus the directional effect of

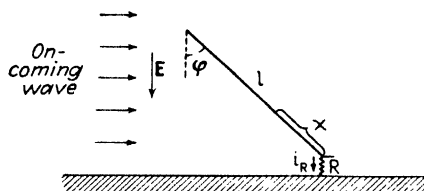



FIG. 16.24.—Tilted-wire antenna.

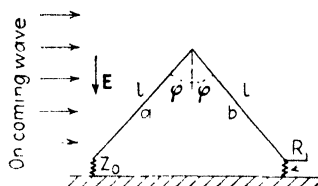



FIG. 16.25.—Folded-wire antenna.

the configuration shown in Fig. 16.25 is more pronounced than that of Fig. 16.24.

The logical development is the diamond-shaped or rhombic antenna which is composed of two folded wires lying in the same plane with a common base, as shown in Fig. 16.26. The analysis is, of course, the same as in the previous case, except that here there are four wires to consider instead of two. Assume that the angle between the incident wave and the plane of the diamond is δ and that the angle between the

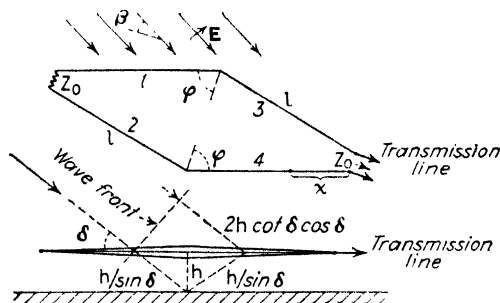



FIG. 16.26.—Horizontal rhombic antenna.

projection of the wave normal on this plane and the longer diagonal of the diamond is β; the electric vector of the wave is taken as lying in the plane of the diamond. The angle of approach is $\varphi + \beta$ for wires 1 and 4 and $\varphi - \beta$ for wires 2 and 3 and the sine of the angle involved in the phase lag is $\sin (\varphi \pm \beta) \cos \delta$, the positive sign referring to wires 1 and 4 and the negative to wires 2 and 3. This general case will not be calculated, but the following expression is easily obtained for a wave along the major axis of the diamond, *i.e.*, $\beta = 0$.

$$\mathbf{i}_R = \frac{jE\lambda}{2\pi \mathbf{z}_0} \frac{\cos\phi}{1 - \cos\delta \sin\phi}[1 - e^{-j\frac{2\pi l}{\lambda}(1 - \cos\delta \sin\phi)}]^2 \qquad (16.47)$$

where $\mathbf{z}_0$ is the characteristic impedance of the antenna wire pairs, which is, of course, equal to the transmission line and terminating impedance. In order to keep this impedance constant along the antenna, it is evident qualitatively from Eq. (14.37) that the size of the wires should increase as their separation increases toward the center of the diamond. This can be accomplished approximately by constructing each side of the diamond of two or three wires which are in contact at the junctions 1-2 and 3-4 and spread apart at the junctions 1-3 and 2-4. In addition there will be in general a wave reflected from the ground which will be assumed to be perfectly reflecting and parallel to the plane of the antenna. The path difference between the reflected and direct waves reaching a point on the antenna is seen from Fig. 16.26 to be $\frac{2h}{\sin\delta} - 2h \cot\delta \cos\delta$ or $2h \sin\delta$. Since the phase change at reflection from a perfect conductor is π, the phase factor for this wave is

$$e^{j\left(\frac{4\pi h}{\lambda}\sin\delta + \pi\right)}.$$

Hence, to get the effect of both waves, Eq. (16.47) must be multiplied by

$$[1 + e^{j\left(\frac{4\pi h}{\lambda}\sin\delta + \pi\right)}] \qquad .$$

On making use of the identity $|1 - e^{jx}| = 2 \sin\frac{x}{2}$ the absolute value of the current in the transmission line to the transmitter or receiver is

$$I_R = \frac{4E\lambda}{\pi Z_0}\left(\frac{\cos\phi}{1 - \cos\delta\sin\phi}\right)\left[\sin\left(\frac{\pi l}{\lambda}(1 - \cos\delta\sin\phi)\right)\right]^2\left[\sin\left(\frac{2\pi h}{\lambda}\sin\delta\right)\right]$$

The optimum values of the antenna constants for the major lobe of the radiation pattern can be obtained by considering the three factors separately. The first factor is a maximum as a function of ϕ when,

$$\sin\phi = \cos\delta$$

The second factor has its first maximum when

$$l = \frac{\lambda}{2(1 - \sin\phi\cos\delta)} = \frac{\lambda}{2\sin^2\delta}$$

The third factor has its maximum for the smallest value of h at

$$h = \frac{\lambda}{4\sin\delta}$$

These three equations permit the design of an antenna which will produce the maximum signal for a particular frequency. If δ is considered as the independent variable, it may be shown that the optimum length

of a side is approximately $\frac{1}{8}$ greater than the value given above. This is the value of l to be used for directing the major lobe at a particular angle. However, in distinction to the standing-wave type of antenna the rhombic type has the great advantage of working well over a considerable range of frequencies around the optimum value. It has the additional advantage that it is less cumbersome and expensive than a comparable array of half-wave antennas.

16.8. The Antenna as a Circuit Element.—The antenna is essentially a circuit element with distributed parameters. As such it is multiply

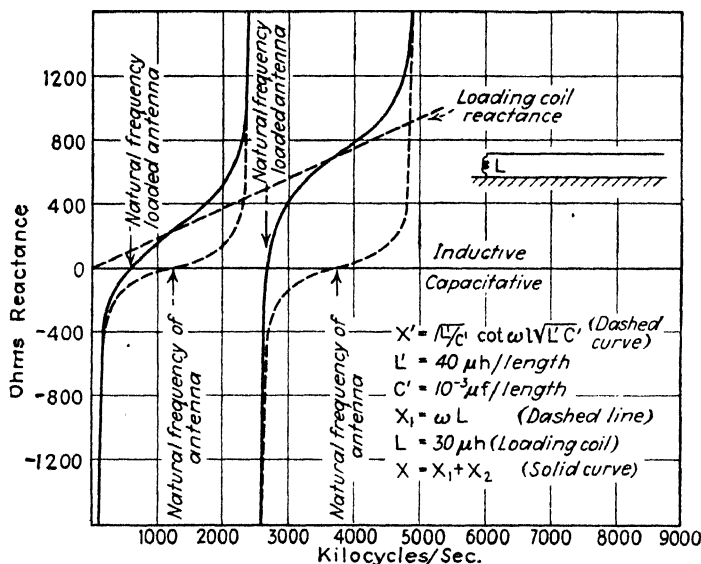



FIG. 16.27.—Long horizontal-wire antenna in series with an inductance.

periodic, and unless the antenna possesses a high degree of symmetry, the natural periods are not even approximately in a harmonic relation to one another. Furthermore, the antenna is generally associated with lumped or localized inductances or capacities in the rest of the circuit, and if the periods of the antenna were originally in a harmonic relation, the presence of these additional elements removes this symmetry. The values of the antenna constants are calculable in only a few relatively simple cases. An illustration of one of these is shown in Fig. 16.27. A long horizontal antenna can be considered as a section of a transmission line formed by itself and its image beneath the surface of the earth. The impedance of such a line is given by Eq. (14.36). If the far end is open and the resistance can be neglected, the reactance reduces to

$$X_a = \left(\frac{L'}{C'}\right)^{\frac{1}{2}} \cot \omega l(L'C')^{\frac{1}{2}}$$

where l is the length of the wire and L' and C' are the inductance and capacitance, respectively, per unit length. The dashed curve of Fig. 16.27 represents this reactance as a function of the frequency for certain assumed values of the parameters. The intersections with the zero axis, *i.e.*, the resonant frequencies, are seen to be equally spaced. If a loading coil, which is a lumped inductance, is placed in series with the antenna, the total reactance is $X_a + \omega L$; this addition is performed graphically, resulting in the solid curves of the figure. From these it is evident that the harmonic relation between the free periods has been lost. The same effect would be produced by the introduction of a capacity or another element with distributed parameters, such as a change in height of the line or the addition of a T termination. Also it is not legitimate to neglect the resistive component of the impedance

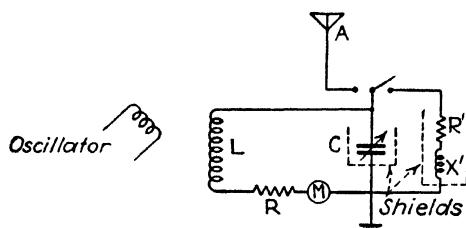



FIG. 16.28.—Schematic circuit for the experimental determination of antenna constants.

of any practical antenna. This further complicates the problem and usually necessitates the experimental determination of antenna constants.

One of the various substitution methods of determining antenna constants experimentally is shown in Fig. 16.28. An absorption type of wavemeter consisting of an inductance, calibrated condenser, and thermal milliammeter is coupled inductively to an oscillator. The wavemeter is set for a particular frequence and the oscillator adjusted to resonance. The coupling is then increased till the meter registers the full-scale deflection. The antenna lead is then connected to the ungrounded side of the condenser and the latter element adjusted for a maximum meter reading and the value of this maximum noted. The nature of this adjustment determines the sign of the reactive component of the antenna, an increase in capacity indicates an inductive and a decrease a capacitative reactance. The antenna lead is then disconnected and a series circuit containing a known resistance and reactance of the proper sign connected across the condenser terminals. These elements R' and X' are then adjusted to the proper values such that the maximum meter reading on varying C is the same as that observed when the antenna was in the circuit. The resistance and reactance of the antenna at the frequency of the oscillator are then equal to the values of R' and X'. The oscillator should have good regulation and be suffi-

ciently powerful to be unaffected by the presence of the measuring circuits. The characteristics of R' and X' must be known at the frequency used and the circuits must be so arranged and shielded that coupling to the oscillator is entirely through the wavemeter circuit itself. By this procedure the impedance of an antenna can be measured over a range of frequencies and its characteristics determined.

The problem of coupling the antenna to a transmission line from a transmitter or receiver at relatively low frequencies involves merely the circuit theory of ordinary lumped constants. The antenna and associated circuit should be tuned for a maximum circulating current. And to avoid reflection of power back to the transmitter the antenna circuit impedance should be matched to that of the line. Two typical circuits for performing these functions are shown in Fig. 16.29. In circuit (*a*) the secondary circuit is tuned by adjusting L_2 or C_2 in the absence of

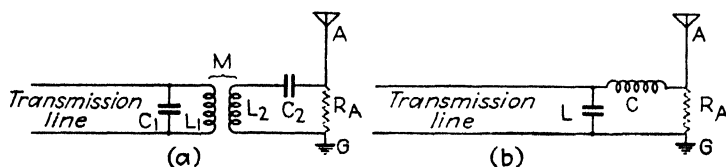



FIG. 16.29.—Typical antenna-coupling circuits.

the primary circuit. The primary is also tuned to the operating frequency and the impedance presented to the line adjusted to the proper value by varying the coupling between the primary and secondary and varying the capacity C_1. Circuit (*b*) is a somewhat simpler one in which the two elements L and C are adjusted to fulfill the required conditions; however, it is possible to satisfy the conditions with this circuit only if the antenna circuit resistance is less than the characteristic resistance of the line.

For high-frequency systems such as half-wave antennas the technique is somewhat different for practical reasons. Frequently the situation of the antenna is such that it is difficult to vary the direct connections to it under operating conditions and at very high frequencies it is difficult to produce satisfactory lumped inductances or capacities. Distributed parameter systems such as resonant or nonresonant lines are generally used. A typical resonant line system is shown in (*a*) of Fig. 16.30. The line and antenna, considered as a single system, are tuned by adjusting the condensers to the frequency for which the antenna is a half-wave doublet. In other words, the condensers are so adjusted that they present an equal but opposite reactance to that presented by the line terminated by the antenna. There are then standing waves on the line, but if the two wires forming it are close together, there is little radiation except from the terminating antenna. An untuned parallel-wire trans-

mission line can be approximately matched in impedance to a half-wave antenna if the characteristic impedance of the line is greater than that of the antenna at its center. For the impedance of a half-wave doublet increases toward the ends and the line spacing may be increased near the antenna and connected symmetrically the proper distance on either side of the mid-point. It is also possible to match the line and antenna impedances by means of a properly spaced line a quarter of a wave length long. Such a line acts exactly like a transformer. This may be seen by considering Eq. (14.36)

$$\mathbf{z}_0 = \mathbf{z}_i \frac{\mathbf{z}_t + \mathbf{z}_i \tanh \gamma l}{\mathbf{z}_i + \mathbf{z}_t \tanh \gamma l} \tag{16.48}$$

where $\mathbf{z}_0$ is the input impedance of the terminated line, $\mathbf{z}_t$ the terminating impedance, $\mathbf{z}_i$ the characteristic impedance of the line, and γ the prop-

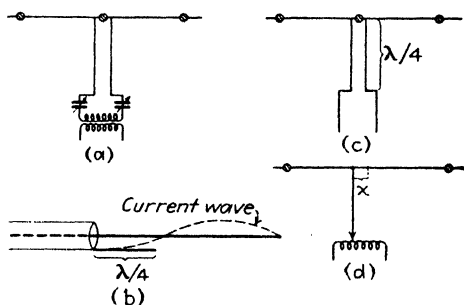



FIG. 16.30.—High-frequency antenna-coupling systems.

agation constant. If $l = \frac{\lambda}{4}$ it is also equal to $v\pi/2\omega$, where v, the wave velocity, is equal to $1/(L'C')^{1/2}$ if the resistance of the line can be neglected. On this same assumption $\gamma = j\omega(L'C')^{1/2}$ and the tanh term becomes $j \tan \frac{\pi}{2}$, which is infinite. Hence the other terms in numerator and denominator can be neglected in comparison and the equation becomes

$$\mathbf{z}_0 = \frac{\mathbf{z}_i^2}{\mathbf{z}_t}$$

Thus the transmission line of impedance $\mathbf{z}_0$ can be matched to the antenna of impedance $\mathbf{z}_t$ by a quarter-wave line of characteristic impedance $(\mathbf{z}_0\mathbf{z}_t)^{1/2}$. Typical instances are illustrated in (*b*) and (*c*) of Fig. 16.30.

It is interesting to analyze approximately a quarter-wave line used in a somewhat different way. This is illustrated in Fig. 16.31. From Eq. (16.48) the impedance $\mathbf{z}_s$ is given by

$$\mathbf{z}_s = \mathbf{z}_i \frac{\mathbf{z}_s' + \mathbf{z}_i \tanh \gamma x}{\mathbf{z}_i + \mathbf{z}_s' \tanh \gamma x}$$

where $\mathbf{z}_i$ is the characteristic impedance of the quarter-wave line and $\mathbf{z}'_s$ is the parallel impedance of $\mathbf{z}_r$ and $\mathbf{z}''_s$, *i.e.*,

$$\frac{1}{\mathbf{z}'_s} = \frac{1}{\mathbf{z}_r} + \frac{1}{\mathbf{z}''_s}$$

And also from Eq. (16.48)

$$\mathbf{z}''_s = \frac{\mathbf{z}_i}{\tanh \gamma\left(\frac{\lambda}{4} - x\right)}$$

since the line is open and the terminating impedance is infinite. Assum-

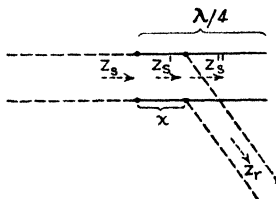



FIG. 16.31.—Diagram for circuit relations in lines coupled through a $\gamma/4$ line.

ing that the resistance of the line can be neglected in comparison with the inductive reactance

$$\gamma = j\omega(L'C')^{1/2} = \frac{2\pi j}{\lambda}$$

On inserting these values in the equation for $\mathbf{z}_s$ it is found that,

$$\mathbf{z}_r = \frac{\mathbf{z}_i^2 \sin^2 \theta}{\mathbf{z}_s} + j\mathbf{z}_i \sin \theta \cos \theta$$

where θ is written for $2\pi x/\lambda$. Thus for suitable values of $\mathbf{z}_i$ and x such a system can be used for matching the impedance between two circuits of impedance $\mathbf{z}_r$ and $\mathbf{z}_s$. If $\cos \theta$ is small or if $\mathbf{z}_i >> \mathbf{z}_s$, the second term can in general be neglected. This has an interesting though approximate relation to the single-wire-feeder system shown in (*d*) of Fig. 16.30. Consider that Fig. 16.31 represents a half-wave antenna parallel to the surface of a perfectly conducting earth and its image antenna below the surface. The impedance $\mathbf{z}_r$ then represents the impedance between the single-wire feeder and the earth. For an impedance match between the wire and antenna from the previous result

$$\mathbf{z}_r = \frac{\mathbf{z}_i^2 \sin^2 \theta}{\mathbf{z}_d} + j\mathbf{z}_i \sin \theta \cos \theta$$

where $\mathbf{z}_d$ is the impedance of a half-wave doublet at the center (Sec. 16.7). Though the resistance of a radiating doublet is not small, the above formula gives quite acceptable results.[1] Except for small values

[1] EVERITT and BYRNE, *Proc. I.R.E.*, **17**, 1840 (1929).

of x the first term alone gives a fair approximation, and even near the center the absolute magnitude of the impedance can be obtained by expanding the expression considering the second term small, yielding

$$Z_r = \frac{R_i^2 \sin^2 \theta}{R_d} + \frac{1}{2} R_d \cos^2 \theta$$

where R_d is the radiation resistance of the doublet and its reactance has been neglected. Though the proper position for the junction of the feeder and antenna can be calculated from this expression, the final adjustment is generally made empirically. When the circuit is properly adjusted, there are no standing waves on the feeder and the current is uniform along it. It will, of course, radiate in the manner described in connection with the inclined-wire antennas, but the radiation is small in comparison with that from the terminating doublet.[1]

Problems

1. The rate of generation of heat in a long cylindrical wire carrying a current i is equal to i^2R where R is the resistance of the wire. Show that this joule heating can be described in terms of Poynting's vector as a flow of energy in from the surrounding space.

2. The solar constant, which is equal to 2.2 cal. per minute per square centimeter, is the mean rate of reception of radiant energy from the sun at noon. What is the rms. value of the electric vector in sunlight at the surface of the earth? What must be its value at the surface of the sun? (Sun's radius $= 7 \times 10^8$ m., distance from sun to earth $= 1.5 \times 10^{11}$ m.)

3. Assuming the solar constant of the previous problem, what is the total radiation force exerted by the sun on the earth?

4. An electromagnetic wave with an effective electric vector of 0.1 volt per meter is incident on an absorbing surface of mass 10^{-3} gm. per square centimeter and specific heat 0.2 cal. per degree centigrade. What is the rate of rise of temperature of the surface?

5. How does the radiation pressure on a perfectly absorbing surface in a beam of light compare with that on a perfect reflector placed in the same beam?

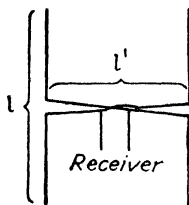
6. A beam of light is incident normally upon a surface of water which has an index of refraction of 1.33. Calculate the electric and magnetic vectors in the reflected and refracted waves and the ratio of the energy reflected to the energy transmitted.

7. Assuming that the radius of curvature of a wave in the ionosphere must be less than that of the earth for it to return, calculate the minimum value of dn/dh for the return of an 80-m. wave, assuming that the ion density is 5×10^5 electrons per cubic centimeter.

8. Taking the average maximum ion density of electrons in the ionosphere as 3×10^6 per cubic centimeter, show that the limiting angle between the initial wave normal and the normal to the surface of the earth for reflection is given approximately by $\cos^{-1} (\lambda/75)$.

9. Find the rms. values of the electric and magnetic vectors in the radiation from a 60-watt lamp at a distance of 1 m., assuming that all the energy supplied to the lamp is radiated.

[1] For the general use of transmission lines as filters and transformers see Mason and Sykes, *Bell, System Tech. J.*, **16**, 275 (1937).

10. A type known as the Adcock antenna is illustrated in the accompanying figure. Show that the emf. generated in the system by the electric vector **E** of an electromagnetic wave is given by $\frac{2ll'E}{\lambda}\cos\theta$, where θ is the angle between **E** and **l** and l' is assumed to be much less than λ. Hence for a maximum signal the sides of the H are in the direction of the electric vector of the wave.

11. A loop antenna has 10 turns and a Q of 75 at 1 megacycle. When it is tuned to this frequency by means of a perfect condenser, what is the ratio of the field strength of the wave to the emf. appearing across the terminals of the condenser?

12. What is the value of the field strength at the surface of a flat perfectly conducting earth 1 km. from a vertical quarter-wave antenna, also on the surface, that carries an effective current of 10 amp. at its base? What is the field strength as observed in an airplane 1 km. above the first point of observation?

13. Instead of rotating a large loop to find the direction of incidence of a radio wave, the following arrangement is sometimes adopted: Two large loops are set up perpendicular to one another with a common vertical diameter. These are then each connected to one of a pair of similarly situated smaller loops within which a third coil is free to rotate about the common diameter. Explain how the direction of incidence of the wave can be determined from the angular setting of the rotating coil for the maximum signal.

14. The effective height of an antenna is defined as the ratio of the emf. generated in it to the field strength of the electromagnetic wave. What is the effective height of (*a*) a vertical wire of length l erected on the surface of the earth, (*b*) a loop of diameter d?

15. Calculate the radiation resistance of a 100-turn coil 10 cm. in diameter at a frequency of 1 megacycle.

16. Show that the power gain in decibels over a single half-wave antenna at the principal maximum obtained from n parallel coplanar half-wave antennas excited in phase is $20 \log_{10} n$. If these are each backed by a suitable antenna to suppress the radiation to the rear, the power gain is $20 \log_{10} 2n$. If the simple half-wave antennas are oriented at random in the plane, the power gain is $10 \log_{10} \frac{n^2}{2}$.

17. Calculate the fraction of the power radiated by a wire short compared to the wave length within 10° of the equatorial plane.

18. Using a table of tangents, calculate the angular zeros of the radiation pattern in the equatorial plane from 10 half-wave antennas which are arranged parallel to one another, spaced $\lambda/2$ apart along a straight line and excited in phase. What is the angular width of the principal maximum?

19. Show that Eq. (16.46) for a folded-wire antenna is the same for a wave incident in the reverse direction except for the sign of $\sin\phi$. Show that if l is an even number of wave lengths, the signal is of the same strength and if l is an odd multiple of $\lambda/4$, there will be no radiation in this reverse direction from the antenna.

20. Calculate the shape, size, and height above the earth for a rhombic antenna to have its major lobe at 10° to its plane for a frequency of 20 megacycles.

21. To a first approximation the resistance of an antenna can be considered as made up of three terms: (1) high-frequency ohmic resistance, (2) dielectric-loss resistance due to neighboring dielectric materials, (3) radiation resistance. Consider a copper wire 1 mm. in diameter and 10 m. long, with an effective total capacity of 10^{-9} farad. From the specific resistance of copper and Prob. 23, Chap. X, calculate and plot the ohmic resistance on a semilogarithmic scale from 10^4 to 10^8 cycles per second. Assuming a constant power factor of 10^{-2}, calculate and plot the dielectric-loss resistance over the same range. Calculate and plot the radiation resistance and add the curves graphically to obtain the total effective resistance of the wire as a function of the frequency.

22. Show that the circuit of Fig. 16.29*b* can only be used if R, the characteristic impedance of the line, is greater than R_a, the antenna resistance, and derive the proper values of L and C in terms of these quantities and the antenna reactance X_a.

23. A quarter-wave two-wire transmission line is used to match the impedance of a 500-ohm line to a 200-ohm antenna. What is the proper spacing between the wires if they are 6 mm. in diameter?

24. From Prob. 28, Chap. IX, show that the magnetic field in the direction θ with the polar angle due to a spinning sphere with a uniform charge upon its surface is $(\mu_0 \omega e a^2/12\pi r^3)\sin\theta$. From the expression for the momentum density $\mathbf{E}\times\mathbf{H}/c^2$ show that the external electromagnetic angular momentum associated with such a sphere is $\mu_0^2\omega e^2 a/18\pi$, and hence the ratio of the magnetic moment to the angular momentum of a spinning electron (if it can be considered to be represented by such a system) is e/m. (From Prob. 20, Chap. IX, the effective radius of an electron is $\mu_0 e^2/6\pi m$.)

25. Extend the argument of Sec. 16.4 regarding the propagation of an electromagnetic wave through an ionized gas to include the assumption that the ion density decreases owing to collisions with gas molecules at a rate given by $N\nu$ per second. The quantity N is the number of ions per unit volume, and ν is proportional to the number of collisions per second of an ion. Taking the current density i as Neu where u is the ion velocity, show that the electric field vector of the wave is related to i in the following way:

$$E = \frac{m}{Ne^2}(\nu + j\omega)i$$

Show from the complex index of refraction that the phase velocity and attenuation constant are given, respectively, by

$$c' = c\left(1 + \frac{\delta}{2}\right), \qquad \alpha = \frac{\delta\nu}{2c}$$

where $\delta = \dfrac{Ne^2}{\kappa_0 m(\nu^2 + \omega^2)}$ is assumed very small in comparison with unity.

26. An electromagnetic wave is propagated through a region of space containing N idealized atoms per unit volume. These atoms possess one electron apiece, which is bound to the heavy atomic nucleus by an elastic force such that the natural angular frequency of oscillation is ω_0 and the logarithmic decrement of free oscillation is δ. Show that the square of the apparent complex index of refraction of the region is given by

$$n^2 = 1 + \frac{Ne^2/\kappa_0 m}{\omega_0^2 - \omega^2 + j\delta\omega\omega_0/\pi}$$

where ω is the angular frequency of the wave. Determine the forms of the propagation and attenuation constants, and plot them as functions of ω/ω_0 for assumed values of $Ne^2/\kappa_0 m$ and δ/π.

27. Show that one of the dominant resonant modes (lowest frequency) of a cubic cavity can be written

$$E'_x = E_0 \sin \frac{\pi y}{A} \sin \frac{\pi z}{A}$$

$$H'_y = +jE_0\sqrt{\frac{\kappa_0}{2\mu_0}} \sin \frac{\pi y}{A} \cos \frac{\pi z}{A}$$

$$H'_z = -jE_0\sqrt{\frac{\kappa_0}{2\mu_0}} \cos \frac{\pi y}{A} \sin \frac{\pi z}{A}$$

where A is the length of one side of the cube. Sketch the nature of the field distributions throughout the cube, and show that $\lambda_0 = \sqrt{2}A$. Also show from Eq. 16.25 that $Q = A/3\delta$ for the dominant mode of a cube, and calculate the numerical value of this for copper.

28. Show by a method analogous to that used in deriving Eq. (16.25) that the ratio of the rate of transportation of energy down a wave guide to its rate of absorption per unit length by the walls is $2fv_g/\omega\delta$. Here v_g is the group velocity of the mode and f is the ratio of the average square value of the magnetic field over the cross sectional area of the guide to the average square value of the magnetic field over the walls.

MATHEMATICAL APPENDIX

A. Taylor's or the General Mean-value Theorem.—A fundamental theorem of mathematics is that if $f(x)$ is a function of x which has derivatives of the first n orders in the interval between $x = a$ and $x = b$, then $f(x)$ may be expressed throughout that interval as

$$f(x) = f(x)_0 + (x - x_0)f'(x)_0 + \frac{(x - x_0)^2}{2!}f''(x)_0 + \frac{(x - x_0)^3}{3!}f'''(x)_0 + \cdots + \frac{(x - x_0)^{(n-1)}}{(n-1)!}f^{(n-1)}(x)_0 + \frac{(x - x_0)^n}{n!}f^n(\xi). \quad (A.1)$$

In this expression x and x_0 lie within the interval, $f(x)_0, f'(x)_0, f''(x)_0$, etc., represent the function and its first, second, etc., derivatives evaluated at $x = x_0$, and in the final term ξ is a quantity whose value lies between x and x_0. This expression is of great value in physical problems because most of the functions encountered satisfy the necessary conditions. A further point of great practical importance is that in many instances only a small range of the variable need be considered, *i.e.*, $(x - x_0)$ is small, or the higher derivatives become vanishingly small so that only a few terms need be retained for a satisfactory representation of the function. When only the second term is retained, it is known as the linear or first-order approximation; when the third term is kept, it gives the quadratic or second-order approximation, etc.

A function of two variables may be handled in an exactly analogous way. If $f(x, y)$ is a function of the two variables x and y, which satisfies for both of these variables the conditions mentioned above, then throughout the prescribed interval

$$f(x, y) = f(x_0, y_0) + \left(h\frac{\partial}{\partial x_0} + k\frac{\partial}{\partial y_0}\right)f(x, y) + \frac{1}{2!}\left(h^2\frac{\partial^2}{\partial x_0} + 2hk\frac{\partial^2}{\partial x_0 \partial y_0} + k^2\frac{\partial^2}{\partial y_0^2}\right)f(x, y) + \frac{1}{(n-1)!}\left(h^{n-1}\frac{\partial^{n-1}}{\partial x_0^{n-1}} + (n-1)h^{n-2}k\frac{\partial^{n-1}}{\partial x_0^{n-1}\partial y_0} + \cdots k^{n-1}\frac{\partial^{n-1}}{\partial y_0^{n-1}}\right)f(x, y) + \frac{1}{n!}\left(h^n\frac{\partial^n}{\partial x_0^n} + nh^{n-1}k\frac{\partial^n}{\partial x_0^{n-1}\partial y_0} + \cdots k^n\frac{\partial^n}{\partial y_0^n}\right)f(x_a, y_b)$$

Here x, y, x_0, and y_0 lie within the necessary interval and $f(x_0, y_0)$ is the particular value of the function at the point $x = x_0$ and $y = y_0$. In the final term $f(x_a, y_b)$ is the value of the function somewhere in the interval between (x, y) and (x_0, y_0). h is an abbreviation for $x - x_0$

and k for $y - y_0$. The partial derivatives are handled just as ordinary algebraic quantities would be. The representative term $\frac{\partial^{n+m}}{\partial x_0^n \partial y_0^m} f(x, y)$ means the nth partial derivative with respect to x and the mth partial derivative with respect to y of the function $f(x, y)$ evaluated at the point $x = x_0$ and $y = y_0$. The extension of the theorem to three or more variables is obvious.

If $f(x, y, z)$ represents the energy of a particle as a function of the coordinates $f(x, y, z) - f(x_0, y_0, z_0)$ represents the increase in energy for the displacement from the point x_0, y_0, z_0 to the point x, y, z. The quantities $-\frac{\partial f}{\partial x_0}, -\frac{\partial f}{\partial y_0}, -\frac{\partial f}{\partial z_0}$ are the first-order force components against which the work is done. If these coefficients vanish, the particle is said to be in equilibrium at the point x_0, y_0, z_0, *i.e.*, there is no force on it and hence no tendency for motion to take place. However, if the quadratic term is negative for an infinitesimal displacement, the displacement will result in a decrease in energy and the equilibrium is unstable. That is, a small displacement will bring into existence forces tending to increase the displacement. However, if this term is positive for all displacements, the equilibrium is stable. The second-order forces are in such a direction as to oppose the displacement and return the particle to the position in which no forces act on it.

B. Fourier Analysis.—It is frequently of the greatest value in many physical problems to be able to expand a function in the form of a trigonometric series. Consider a function $f(t)$ which is defined for the range of the variable from $t = -\pi$ to $t = +\pi$. If this range can be broken up into a finite number of intervals within each of which the function is bounded, continuous, and monotonic, the function can be written as the sum of a series of sine and cosine terms as follows:

$$f(t) = \frac{a_0}{2} + a_1 \cos t + a_2 \cos 2t + a_3 \cos 3t \cdots a_n \cos nt \cdots$$
$$+ b_1 \sin t + b_2 \sin 2t + b_3 \sin 3t \cdots b_n \sin nt \cdots$$

or

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{n=\infty} (a_n \cos nt + b_n \sin nt) \qquad (B.1)$$

where

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \cos nt\, dt$$
$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \sin nt\, dt$$

If $f(t)$ is continuous, the series represents the function accurately; at a point of discontinuity, but of limited fluctuation of $f(t)$, the series represents the mean value of the function as the point of discontinuity is approached from the two sides. A function $f(x)$, which is defined over the range d rather than over the range 2π, may be expanded in a series of this type by changing the variable in the above representation from t to $2\pi x/d$.

Frequently it is not an analytic function but an empirical curve or set of experimental points which is to be represented by means of a trigonometric series. Assume, for example, that the range is 2π and that there are $2p$ experimental points equally spaced along the axis of the variable. If the ordinates of the points are A_1, A_2, A_3, etc., the coordinates specifying them are then

$$\left(\frac{\pi}{p}, A_1\right), \quad \left(\frac{2\pi}{p}, A_2\right), \quad \cdots \quad \left(\frac{q\pi}{p}, A_q\right) \quad \cdots \quad (2\pi, A_{2p})$$

The series to be fitted to these points is

$$f(t) = \frac{a_0}{2} + a_1 \cos t + a_2 \cos 2t + \cdots a_p \cos pt + b_1 \sin t + b_2 \sin 2t + \cdots b_p \sin pt$$

There appear to be $2p + 1$ coefficients, but the last sine term is zero so that the coefficient is not significant. It may be shown that the series represents a function that passes through all the points if the coefficients are given by the following expressions:

$$a_n = \frac{2}{2p+1} \sum_{q=1}^{q=2p} A_q \cos \frac{qn\pi}{p}$$

$$b_n = \frac{2}{2p+1} \sum_{q=1}^{q=2p} A_q \sin \frac{qn\pi}{p}$$

All the cosine terms in a_0 are, of course, unity. The resultant series is the simplest trigonometric series that passes through all the chosen points. The behavior of the function has, of course, not been specified between these points. However, unless the points represent a curve with unusual features, the trigonometric representation is generally adequate for most purposes if about 20 points are fitted in the range. This carries the series up to the tenth harmonic.

C. Elementary Differential Equations. *General Considerations.*—One of the most common equations encountered in the study of elementary physical phenomena is the homogeneous linear differential equation of the second order with constant coefficients

$$A\frac{d^2x}{dt^2} + B\frac{dx}{dt} + Cx = f(t) \qquad (C.1)$$

Here A, B, and C are constants and $f(t)$ is any function of t. Before a complete discussion of this equation is possible, it is necessary to consider the auxiliary equation

$$A\frac{d^2x}{dt^2} + B\frac{dx}{dt} + Cx = 0 \qquad (C.2)$$

This is the same as Eq. (C.1), except that 0 has replaced $f(t)$. If x_1 and x_2 are any two solutions of Eq. (C.2), *i.e.*, if when either x_1 or x_2 are substituted for x in this equation, the left side becomes equal to zero, then $x_1 + x_2$ is also a solution. This may readily be seen by substitution. A third solution, x_3, may be added to the sum of x_1 and x_2 and the result is also a solution. So the sum of any number of solutions is a solution. Similarly, a constant times any solution is a solution. These statements are true, as may easily be seen, for a homogeneous linear differential equation of any order for which the right side is zero.

The general solution of either of these equations is a function of t containing two arbitrary constants. If the differential equation is of the first order, only one arbitrary constant need appear in the general solution. Both first- and second-order differential equations of this general type will be considered in subsequent sections of this Appendix. In deriving the solutions of these equations [Eqs. (C.5) and (C.1′)] the necessity for the introduction of 1 and 2 arbitrary constants, respectively, will be shown. Similarly, it may be proved that if the equation were of the rth order, r arbitrary constants would appear in the general solution. There constants can be evaluated only from auxiliary conditions, such as the values of the solution or its derivatives for particular values of the variable. Each auxiliary condition permits the elimination of one arbitrary constant. Thus only one condition is necessary to eliminate the arbitrary constant occurring in the general solution of a first-order differential equation, but two conditions must be used to eliminate these constants from the solution of an equation of the second order. It is also obvious that to any particular integral of Eq. (C.1) may be added the general solution of Eq. (C.2); the result is also a solution of Eq. (C.1). For this is simply equivalent to adding zero to both sides of this equation.

First-order Equations.—Before proceeding to the solution of Eq. (C.1), some simpler equations obtained by giving particular values to the constants A, B, and C and to the function $f(t)$ will be considered. First, let A and $f(t)$ be zero, in which case the equation reduces to

$$B\frac{dx}{dt} + Cx = 0 \qquad (C.3)$$

This type of equation is encountered, for instance, in resistance-capacity

and resistance-inductance circuits. It may be solved by separating the variables

$$\frac{dx}{x} = -\frac{C}{B}dt = -g\,dt$$

where g is written for C/B. This may be integrated immediately, yielding

$$\begin{aligned}\log_e x &= -gt + \text{const.} \\ &= -gt + \log_e \alpha\end{aligned}$$

or

$$x = \alpha e^{-gt} \tag{C.4}$$

As the equation is of the first order, α is the only arbitrary constant appearing. This constant may be determined in terms of the value of x at a particular value of t; for example, if $x = x_0$ when $t = 0$, then $\alpha = x_0$ and Eq. (C.4) becomes $x = x_0 e^{-gt}$.

If A is zero but $f(t)$ is not, the equation is of the form

$$B\frac{dx}{dt} + Cx = f(t) \tag{C.5}$$

On multiplying through by e^{gt} and dividing by B the equation may be integrated directly

$$e^{gt}\frac{dx}{dt} + ge^{gt}x = \frac{e^{gt}}{B}f(t)$$

or

$$\frac{d(xe^{gt})}{dt} = \frac{e^{gt}}{B}f(t)$$

and integrating

$$xe^{gt} = \frac{1}{B}\int e^{gt}f(t)dt + \text{const.}$$

Putting the constant equal to α, the general solution is found to be

$$x = \alpha e^{-gt} + \frac{e^{-gt}}{B}\int e^{gt}f(t)dt \tag{C.6}$$

Again there is only one arbitrary constant as the equation is of the first order. In order to determine α in terms of the value of x for a particular value of t the form of the function $f(t)$ must be known.

Second-order Equations.—If B and $f(t)$ are both zero, the equation of free, undamped simple harmonic motion results. It is also the equation representing the behavior of inductance-capacity circuits when the resistance is assumed to be negligible.

$$A\frac{d^2x}{dt^2} + Cx = 0 \tag{C.7}$$

It is of interest to consider this by a special method. Let $dx/dt = y$

$$\frac{d^2x}{dt^2} = \frac{dy}{dt} = \frac{dy}{dx}\frac{dx}{dt} = y\frac{dy}{dx}$$

Making this substitution and multiplying through by dx

$$Ay\,dy + Cx\,dx = 0$$

Integrating

$$A\frac{y^2}{2} + C\frac{x^2}{2} = E \tag{C.8}$$

where E is the constant of integration. If A and $1/C$ are the inductance and capacity, and x and y the charge and current, respectively, the left side of Eq. (C.8) represents the sum of the magnetic and electric energies stored in a circuit. Thus the first integral of Eq. (C.7) expresses the conservation of energy in this idealized circuit. To complete the solution of the original equation, y is rewritten in terms of dx/dt, and Eq. (C.8) becomes on rearranging

$$\frac{dx}{dt} = \sqrt{\frac{C}{A}}\sqrt{\frac{2E}{C} - x^2}$$

The variables are separable and the resulting equation may be integrated directly

$$\int \frac{dx}{\sqrt{\frac{2E}{C} - x^2}} = \sqrt{\frac{C}{A}}\int dt$$

$$\sin^{-1}\sqrt{\frac{C}{2E}}\,x = \sqrt{\frac{C}{A}}\,t + \delta$$

$$x = \sqrt{\frac{2E}{C}}\sin\left(\sqrt{\frac{C}{A}}\,t + \delta\right) \tag{C.9}$$

Here δ, the second constant of integration, is known as the *phase angle*. The energy E and the phase angle δ are determined from the initial conditions.

The more general second-order equation Eq. (C.2) will next be considered

$$A\frac{d^2x}{dt^2} + B\frac{dx}{dt} + Cx = 0$$

The simplest method of handling this equation is to assume that $x = e^{kt}$

is a particular solution. Substituting this expression in Eq. (C.2) and dividing by e^{kt}, the equation reduces to

$$Ak^2 + Bk + C = 0$$

If k satisfies this equation, the assumed form is a solution of Eq. (C.2). Solving for k

$$\begin{aligned} k &= -\frac{B}{2A} \pm \sqrt{\frac{B^2}{4A^2} - \frac{C}{A}} \\ &= -a \pm m \end{aligned}$$

where the abbreviations $a = B/2A$, and $m = \sqrt{a^2 - C/A}$ have been used. There are two possible values for k given by the choice of signs in front of m. On multiplying each of these particular solutions by an arbitrary constant and adding them together, a solution containing two arbitrary constants, which is therefore the general solution of the equation, is obtained.

$$x = (\alpha e^{mt} + \beta e^{-mt})e^{-at} \tag{C.10}$$

Here α and β are the two arbitrary constants. This expression may be put into three special forms, depending on the value of m. They are of sufficient importance to be discussed separately.

Case 1. $B^2 > 4AC$, which corresponds to m being real: It is generally more convenient to express α and β in terms of the initial conditions. Let $x = x_0$ and $dx/dt = x'$ when $t = 0$. Putting these values in Eq. (C.10) and its derivative with respect to t yields the two equations

$$\begin{aligned} x_0 &= \alpha + \beta \\ x' &= m(\alpha - \beta) - a(\alpha + \beta) \end{aligned}$$

Solving these for α and β

$$\alpha = \frac{1}{2m}[(m + a)x_0 + x']$$

$$\beta = \frac{1}{2m}[(m - a)x_0 - x']$$

Inserting these values in Eq. (C.10)

$$x = \frac{e^{-at}}{m}\left[(x' + ax_0)\left(\frac{e^{mt} - e^{-mt}}{2}\right) + x_0 m\left(\frac{e^{mt} + e^{-mt}}{2}\right)\right] \tag{C.11}$$

The first bracket containing the exponentials is by definition the hyperbolic sine function and the second is the hyperbolic cosine function. Sinh mt and cosh mt could be written for these brackets at this stage,

but a further simplification can be made. A quantity δ may be defined by the equations

$$x' + ax_0 = Q \cosh \delta \qquad \text{and} \qquad mx_0 = Q \sinh \delta$$

Since $\cosh^2 \delta - \sinh^2 \delta = 1$, the following condition is imposed upon Q: $Q = \sqrt{(a^2 - m^2)x_0^2 + 2ax_0x' + x'^2}$. Since the ratio of the hyperbolic sine to the hyperbolic cosine is the hyperbolic tangent, δ may be written as

$$\delta = \tanh^{-1}\left(\frac{mx_0}{x' + ax_0}\right)$$

Using Q and δ, Eq. (C.11) becomes

$$x = \frac{e^{-at}Q}{m}(\cosh \delta \sinh mt + \sinh \delta \cosh mt)$$

or since the bracket is equal to $\sinh (mt + \delta)$

$$x = \frac{e^{-at}Q}{m} \sinh (mt + \delta) \qquad (C.12)$$

This is the most compact and convenient form for the so-called *aperiodic* or *nonoscillatory-solution* of Eq. (C.2). Sinh mt increases steadily with t, but not so rapidly as e^{at} so that x approaches zero as t becomes very large. There may be one maximum value for the function but no more.

Case 2. $B^2 < 4AC$, which corresponds to m being imaginary: It is here more convenient to write $\sqrt{-1}$ explicitly as j and define a real quantity, m', by $m = jm'$. Putting jm' for m in Eq. (C.11), it may be reduced in an exactly analogous way in terms of the circular functions sine and cosine. Using the identity $\sin^2 \delta' + \cos^2 \delta' = 1$ and defining the new quantity δ' by

$$\delta' = \tan^{-1}\left(\frac{m'x_0}{x' + ax_0}\right)$$

the solution of the equation becomes

$$x = \frac{e^{-at}Q}{m'} \sin (m't + \delta') \qquad (C.13)$$

This is the most convenient form of the so-called *damped oscillatory solution* of Eq. (C.2) in terms of the initial conditions. The sine function oscillates between plus and minus one, but its amplitude is continuously decreased by the exponential factor. By differentiating Eq. (C.13), the maxima are seen to occur at the times determined by the roots of the equation

$$\tan (m't + \delta') = \frac{m'}{a}$$

This is periodic with the period $\tau = 2\pi/m'$ as is Eq. (C.13) itself. The ratio of successive maximum amplitudes is thus $e^{-a\tau}$ from Eq. (C.13). The exponent $\tau a = 2\pi a/m'$ is frequently written δ and is known as the logarithmic decrement.

Case 3. $B^2 = 4AC$ which corresponds to $m = 0$: If m is set equal to zero in Eq. (C.11), x becomes indeterminate. This may be avoided by expanding the exponential functions in the first bracket in accordance with the formula

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

If m is now allowed to approach zero, the solution becomes

$$x = [x_0 + (x' + ax_0)t]e^{-at} \qquad (C.15)$$

This is the so-called *critically damped* solution. As t increases less rapidly than the hyperbolic sine, x decreases more rapidly for critical damping than for the general aperiodic case.

Proceeding now to the solution of the general equation of the type of Eq. (C.1), it is here rewritten for convenience with the right side multiplied by A.

$$A\frac{d^2x}{dt^2} + B\frac{dx}{dt} + Cx = Af(t) \qquad (C.1')$$

The quantity $\sqrt{-1}$, written as j, is necessary for the representation of the square root of a negative number. Since the squares of all ordinary or "real" numbers are positive, the numbers from which j may be factored to obtain a real number are of a different character and are known as "imaginary" numbers. They are written in the form jA, where A is real. Sums of real and imaginary numbers, which are known as "complex numbers," may be represented geometrically by erecting an axis perpendicular to that of the real numbers along which imaginary numbers may be considered to lie. The complex number $x + jy$ is represented by a point in the plane determined by these axes which has an abscissa of x and an ordinate of y. Since from Fig. C.1 $x = r \cos \varphi$ and $y = r \sin \varphi$, an alternative representation of the point is $r(\cos \varphi + j \sin \varphi)$. If the sine and cosine functions are written in terms of the exponential function, the bracket reduces to $e^{j\varphi}$. Thus the general complex number **z** will be written in boldface type and may be expressed in any of the following forms:

FIG. C.1.—Representation of a complex number.

$$\mathbf{z} = x + jy = r(\cos \varphi + j \sin \varphi) = re^{j\varphi} \qquad (C.14)$$

where

$$r = \sqrt{x^2 + y^2} \quad \text{and} \quad \varphi = \tan^{-1}\frac{y}{x}$$

A simplification is introduced by writing $x = ze^{k_1t}$, where z is a function of t and k_1 an abbreviation for $m - a$. Inserting this expression for x in Eq. ($C.1'$) and substituting for A, B, and C in terms of a and m, the equation reduces to the following:

$$\frac{d^2z}{dt^2} + 2m\frac{dz}{dt} = e^{-k_1t}f(t)$$

This may be integrated once immediately, yielding

$$\frac{dz}{dt} + 2mz = \int e^{-k_1t}f(t)\,dt + 2m\beta$$

where $2m\beta$ is the constant of integration. This may now be integrated again in the manner of Eq ($C.5$) to give

$$z = \alpha e^{-2mt} + e^{-2mt}\{\int e^{2mt}[\int e^{-k_1t}f(t)dt + 2m\beta]dt\}$$

Multiplying this by e^{k_1t} to obtain x and using the abbreviation

$$k_2 = (m + a),$$

there results with some rearrangement

$$x = \alpha e^{k_2t} + e^{k_2t}\{\int e^{2mt}[\int e^{-k_1t}f(t)dt]dt + 2m\beta\int e^{2mt}\,dt\}$$

Performing the last integration and taking an e^{-at} out

$$x = e^{-at}(\alpha e^{-mt} + \beta e^{mt} + e^{-mt}\{\int e^{2mt}[\int e^{-k_1t}f(t)dt]dt\})$$

Integrating the last term by parts and collecting terms, the final result becomes

$$x = e^{-at}\left\{\alpha e^{-mt} + \beta e^{mt} + \frac{1}{2m}\left[e^{mt}\int e^{-k_1t}f(t)dt - e^{-mt}\int e^{-k_2t}f(t)dt\right]\right\} \quad (C.16)$$

This is the general solution of Eq. ($C.1'$). α and β are the two arbitrary constants which must appear; they cannot be interpreted in terms of the initial conditions without a knowledge of $f(t)$.

D. General Vector Theory Necessary for the Description of Elementary Electric and Magnetic Phenomena.—Many familiar quantities such as mass, temperature, charge, pole strength, etc., are completely specified when a single number representing their magnitude is given. These quantities are known as *scalars*. In addition to these there are other quantities such as velocity, force, current, electric- and magnetic-field strength, etc., which require for their specification not only a magnitude but a direction in space. These quantities are known as *vectors*. A vector may be represented geometrically by an arrow; the length of the arrow represents on some arbitrary scale the magnitude of the

vector, and the direction associated with the vector is that in which the arrow points. It should be mentioned that all quantities capable of representation by an arrow are not vectors in the sense in which the term is here used. All directed electric and magnetic quantities are vectors and obey the laws of manipulation which are given below. There exist, however, other quantities such as rotations of a rigid body which may also be represented by an arrow but which do not obey these laws and hence are not considered as vectors. Directed quantities which are not vectors do not arise in the study of electricity and magnetism. Throughout this book scalar quantities are written in ordinary type and vector quantities in boldface type.

Sum of Vectors.—Two vectors are said to be equal if they are equal both in direction and magnitude. The process of addition of vectors is defined as follows: The initial point of vector **B** is placed at the terminal point of vector **A**. **C**, which is the sum of these two, is then the vector joining the initial point of **A** with the terminal point of **B**. This is the familiar parallelogram law of composition. It is written algebraically in the same form as the addition of scalars

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

It may be seen from the geometrical construction that the process of addition of vectors obeys the familiar commutative and associative laws of the addition of scalars, *i.e.*,

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

and

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) = \mathbf{B} + (\mathbf{A} + \mathbf{C})$$

The sum of two equal vectors **A** and **A** is a vector in the direction of **A** but with twice the magnitude of **A**, *i.e.*, the vector 2**A**. Thus the product of a scalar n and a vector **A**, which is written $n\mathbf{A}$, is a vector in the direction of **A** with a length equal to n times that of **A**. If **a** is a vector of length equal to unity on the scale chosen, $n\mathbf{a}$ is a vector of length n in the direction determined by **a**; n is numerically equal to the magnitude of the vector.

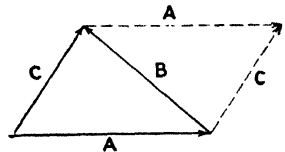



FIG. *D*.1.—Addition of vectors.

The vector $-\mathbf{A}$ is equal in magnitude to **A** but directed in the opposite sense. The process of subtraction is the opposite of that of addition. For instance, in Fig. *D*.1, $\mathbf{B} = \mathbf{C} - \mathbf{A}$.

Vectors may be conveniently represented in Cartesian coordinates by means of the unit vector concept. If Ox and Oy in Fig. *D*.2 are the positive directions of these two axes, the positive direction of the third axis can evidently be chosen in two ways. In the later development of

vector analysis this choice is significant. The positive direction of the axis Oz should be so chosen that for an observer looking along this positive direction a clockwise rotation about this axis through a right angle brings Ox into the position previously occupied by Oy. These axes are then known as a right-handed coordinate system. Let the vectors **i**, **j**, and **k** be vectors of unit length along the x, y, and z axes, respectively.

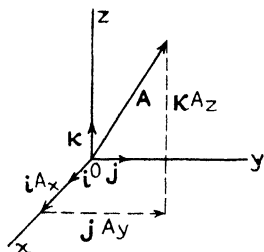



FIG. D.2.—Resolution of a vector.

A vector **A** can be considered to be the sum of three vectors, each one of which lies along one of these axes. Writing these as in Fig. D.2 as the products of the scalar quantities A_x, A_y, and A_z and the unit vectors **i**, **j**, and **k**

$$\mathbf{A} = \mathbf{i}A_x + \mathbf{j}A_y + \mathbf{k}A_z$$

A_x, A_y, and A_z are known as the Cartesian components of the vector **A**. The components of the sum of two vectors are evidently the sums of the corresponding components of the two vectors, *i.e.*,

$$\mathbf{A} + \mathbf{B} = \mathbf{i}(A_x + B_x) + \mathbf{j}(A_y + B_y) + \mathbf{k}(A_z + B_z)$$

Scalar Product.—In many physical problems a quantity of great importance is the product of the magnitude of one vector by the magnitude of the projection of another vector upon it. This quantity is a scalar and possesses the characteristics of an ordinary algebraic product and hence is called the *scalar product*. It is written $\mathbf{A} \cdot \mathbf{B}$ and from its definition

$$\mathbf{A} \cdot \mathbf{B} = AB \cos (\mathbf{AB}) \tag{D.1}$$

Here A and B are the magnitudes of the two vectors and $(\mathbf{AB})$ is the angle included between them. From the definition of the scalar product it is evident that it resembles the ordinary algebraic product in being commutative and distributive, *i.e.*,

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$$

and

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C} + \mathbf{D} + \cdots) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C} + \mathbf{A} \cdot \mathbf{D} + \cdots$$

If the angle $(\mathbf{AB})$ is acute, the product is positive; if it is obtuse, the product is negative. If the vectors are parallel, the scalar product is simply the product of the magnitudes and if the vectors are perpendicular, it is zero. The scalar product of two equal vectors $\mathbf{A} \cdot \mathbf{A} = \mathbf{A}^2$ may be written as A^2.

From the above definition the scalar products of the unit Cartesian vectors evidently obey the relations

$$\mathbf{i} \cdot \mathbf{j} = \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{i} = 0$$

and

$$\mathbf{i} \cdot \mathbf{i} = \mathbf{j} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{k} = 1 \qquad (D.2)$$

Utilizing the above relations, the scalar product of two vectors may be written in terms of their components. Expressing **A** and **B** in terms of their Cartesian components

$$\mathbf{A} = \mathbf{i}A_x + \mathbf{j}A_y + \mathbf{k}A_z \qquad \text{and} \qquad \mathbf{B} = \mathbf{i}B_x + \mathbf{j}B_y + \mathbf{k}B_z$$

multiplying out in the usual way and then substituting the values of the products of the unit vectors the scalar product becomes

$$\mathbf{A} \cdot \mathbf{B} = A_xB_x + A_yB_y + A_zB_z \qquad (D.3)$$

FIG. D.3.—Projection of the vector **F** on the vector **D** for discussing the scalar product.

Thus the scalar product of two vectors is the sum of the products of the corresponding components. In the case of three or more vectors the order in which the products are formed must be specified. Products occurring within parentheses are to be formed first.

$$\mathbf{A}(\mathbf{B} \cdot \mathbf{C}) \neq \mathbf{B}(\mathbf{C} \cdot \mathbf{A}) \neq \mathbf{C}(\mathbf{A} \cdot \mathbf{B})$$

i.e., the associative law of ordinary scalar quantities is not obeyed by the scalar product. These three vectors are in the directions of **A, B,** and **C,** respectively. The product of the vector **A** and the scalar $\mathbf{B} \cdot \mathbf{C}$ is a vector in the direction of **A** with a magnitude ABC cos (**BC**).

The value of the concept of the scalar product is evident from the following examples: If **F** in Fig. *D*.3 represents a force acting on a body and **D** a displacement of the body, then the work W done on the body during the displacement is by definition FD cos (**FD**). Thus

$$W = FD \cos (\mathbf{FD}) = \mathbf{F} \cdot \mathbf{D}$$

If **F** is a function of position and **dl** is an infinitesimal vector representing the displacement in the neighborhood of a point, the infinitesimal increment of work dW done on the body in the neighborhood of the point is given by: $dW = \mathbf{F} \cdot \mathbf{dl}$. The scalar product has another useful geometrical interpretation. A plane area may be represented by a vector perpendicular to the area and of length equal to the magnitude of the area. If **S** is a vector representing the shaded area of Fig. *D*.4 and **v** is a vector equal to a generator of the cylinder, the volume V of the solid figure swept out by displacing the area by amount **v** is Sv cos (**Sv**), or $V = \mathbf{S} \cdot \mathbf{v}$. If **ds** is an element of area and **v** represents the velocity of an incompressible fluid, the quantity of fluid crossing **ds** in unit time is $\mathbf{v} \cdot \mathbf{ds}$. If **ds** is an element of a closed surface, it is by convention

chosen to be in the direction of the outward normal so that $\mathbf{v} \cdot \mathbf{ds}$ is the outward flow through the element.

Vector Product.—In addition to the scalar product there is another product-like quantity which is of great importance in physical problems. It is known as the *vector product.* The vector product of two vectors **A** and **B** is written $\mathbf{A} \times \mathbf{B}$ and is defined as a vector perpendicular to the plane determined by **A** and **B** and equal in magnitude to the area of the parallelogram whose sides are **A** and **B**. This area is AB sin (**AB**). Furthermore, $\mathbf{A} \times \mathbf{B}$ is directed in such a sense that for an observer sighting in the direction of the arrow a clockwise rotation about the axis $\mathbf{A} \times \mathbf{B}$ through the smaller angle brings **A** into the position occupied

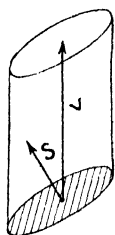



FIG. *D*.4.—The vector **S**, perpendicular to the shaded surface, represents it in direction and magnitude.

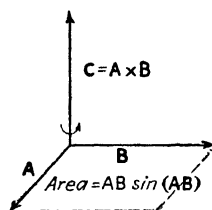



FIG. *D*.5.—Illustration of the vector product.

by **B**. This is the sense of rotation associated with a right-hand screw. Thus, if **C** is the vector product of **A** and **B**, it is directed as shown in Fig. *D*.5 and its magnitude is given by:

$$C = AB \sin (\mathbf{AB}) \qquad (D.4)$$

It is a maximum, *i.e.*, AB, when **A** is perpendicular to **B**, and zero when **A** is parallel to **B**. From the definition it is evident that the vector product is not commutative but

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$$

It may be shown, however, that it does obey the distributive law

$$\mathbf{A} \times (\mathbf{B} + \mathbf{C} + \mathbf{D} + \cdots) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C} + \mathbf{A} \times \mathbf{D} + \cdots$$

The vector products of the unit Cartesian vectors are seen to obey the following relations:

$$\begin{aligned} \mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} &= \mathbf{k} \times \mathbf{k} = 0 \\ \mathbf{i} \times \mathbf{j} &= -\mathbf{j} \times \mathbf{i} = \mathbf{k} \\ \mathbf{j} \times \mathbf{k} &= -\mathbf{k} \times \mathbf{j} = \mathbf{i} \\ \mathbf{k} \times \mathbf{i} &= -\mathbf{i} \times \mathbf{k} = \mathbf{j} \end{aligned} \qquad (D.5)$$

The vector product of **A** and **B** may be expressed in terms of their components and the unit vectors. Expressing **A** and **B** in terms of these unit vectors, multiplying out in the usual way, and then substituting the above expressions for the unit products

$$\begin{aligned}\mathbf{A} \times \mathbf{B} &= (\mathbf{i}A_x + \mathbf{j}A_y + \mathbf{k}A_z) \times (\mathbf{i}B_x + \mathbf{j}B_y + \mathbf{k}B_z) \\ &= \mathbf{i}(A_yB_z - A_zB_y) + \mathbf{j}(A_zB_x - A_xB_z) + \mathbf{k}(A_xB_y - A_yB_x) \quad (D.6)\end{aligned}$$

This sum is the same as the expansion of a determinant; hence the vector product may be written alternatively as

$$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \qquad (D.6')$$

Since **A** × **B** is a vector perpendicular to **A** and **B** and equal in magnitude to the area of the parallelogram formed by them, the scalar product of a vector **C** with this vector product is equal to the volume of the parallelepiped whose edges are these three vectors. This may be seen from Fig. *D*.6. It is evident that the product symbols may be interchanged without altering the value of the triple product. Of course, the vector product must be formed first but with this convention: $\mathbf{A} \times \mathbf{B} \cdot \mathbf{C} = \mathbf{A} \cdot \mathbf{B} \times \mathbf{C}$. Hence this scalar-vector triple product may be written simply **ABC**. A further consideration shows that a cyclic permutation of the factors does not alter the value of the product, but a noncyclic interchange alters the sign.

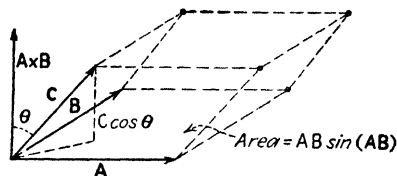



Fig. *D*.6.—Parallelogram whose volume is represented by **ABC**.

$$\mathbf{ABC} = \mathbf{BCA} = \mathbf{CAB} = -\mathbf{ACB} = -\mathbf{CBA} = -\mathbf{BAC}$$

If this triple product is expanded in terms of Cartesian components, it is seen to be identical with the expansion of the following determinant:

$$\mathbf{ABC} = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}$$

The triple vector product also occurs frequently and it can be shown by expansion that it is expressible in terms of two scalar products

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B}) \qquad (D.7)$$

The concept of the vector product is very useful in many physical problems. If a fixed point in a body is chosen as origin and a vector **r**

joins this point with that at which a force **F** is applied the torque **T** tending to rotate the body is given by $\mathbf{T} = \mathbf{r} \times \mathbf{F}$. For F sin (**rF**) is the component of **F** perpendicular to **r**, hence the magnitude of **T** is rF sin (**rF**) and **T**, which is perpendicular to **r** and **F**, is along the axis of rotation.

Differentiation and Integration of Vectors.—Since differentiation with respect to a scalar variable is merely the limit approached in a process of subtraction followed by division by a scalar, and since these processes obey the ordinary laws of scalar manipulation, the concept of a derivative is easily extended to vectors. Thus if $\mathbf{F}(u)$ is a vector function of the scalar variable u, the derivative of **F** with respect to u is defined as

$$\frac{d\mathbf{F}}{du} = (\lim \Delta u \rightarrow 0)\frac{\mathbf{F}(u + \Delta u) - \mathbf{F}(u)}{\Delta u}$$

If $\Delta\mathbf{F}$ is written for the numerator of this fraction, the derivative is a vector in the direction of $\Delta\mathbf{F}$ as Δu approaches zero and equal in magnitude to $\Delta\mathbf{F}/\Delta u$ at this limit. From this definition it is evident that the derivative of the sum of two vectors is the sum of the derivatives and that the derivative of a product is the same as that for the derivative of a product of scalars with, of course, due regard for the change in sign if the order of the factors is changed in the case of the vector product.

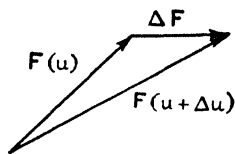



Fig. D.7.—Sum of a finite and infinitesimal vector illustrating vector differentiation.

Similarly, the process of integration is merely the limit approached in a summation of simple products so that it is formally the same for vectors as for scalars. If, for example, **F** represents the force per unit volume on a solid body (for instance, gravitation) it will in general vary from point to point within the body and the total or resultant force acting on the body will be the triple integral: $\iiint \mathbf{F}\, dx\, dy\, dz$ over the volume occupied by the body. Abbreviating the volume element $dx\, dy\, dz$ as dv, this integral may be written more briefly as

$$\int_V \mathbf{F}\, dv$$

Such a vector function as **F** which has values throughout a region constitutes what is known as a *vector field* just as a scalar function of position which has values throughout a region is known as a *scalar field.* A particularly useful integral concept is that of a *line integral.* The line integral of a vector **F** along a curve S from the point A to the point B is defined as the limit for **dl** infinitesimal of the sum of the scalar products $\mathbf{F} \cdot d\mathbf{l}$ along this curve from the point A to the point B. It is written

$$\int_{A(s)}^{B} \mathbf{F} \cdot d\mathbf{l}$$

Its value depends in general on the particular path chosen, *i.e.*, the value of the integral is in general different for the path S' than for the path S. The line integral around a completely closed path, for example, from A to B along S and from B to A along S' in Fig. $D.8$, is written

$$\oint \mathbf{F} \cdot \mathbf{dl}$$

Of course, $\int_A^B \mathbf{dl}$ is simply the vector from A to B and is independent of the path. Another useful concept is the integral of a scalar product over a surface bounded by a closed curve. From the previous discussion in connection with the scalar product, if $\mathbf{v}$ is the velocity of an incompressible fluid and $\mathbf{ds}$ is an infinitesimal vector in the direction of the normal (outward if the surface is closed) to the surface, then $\int \mathbf{v} \cdot \mathbf{ds}$ represents the total flow through the surface of integration per unit time. If the integration is carried out over a closed surface, it may be written

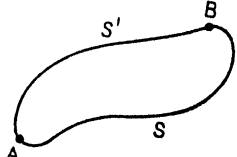



FIG. $D.8$.—Paths of integration.

$$\oint \mathbf{v} \cdot \mathbf{ds}$$

and this quantity represents the total outward flow per unit time from the bounded volume.

Gradient of a Scalar Field.—If a scalar quantity such as temperature, density, potential energy, etc., has a uniquely determined value at each point in a region, it constitutes what has been called a *scalar field*. If u, which is in general a function of x, y, and z, represents this quantity at a point and $u + du$ its value at a point an infinitesimal distance away then by the first-order approximation of Taylor's theorem

$$du = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy + \frac{\partial u}{\partial z}dz$$

The infinitesimal vector displacement $\mathbf{dl}$ may be written as

$$\mathbf{dl} = \mathbf{i}\,dx + \mathbf{j}\,dy + \mathbf{k}\,dz$$

and a new quantity known as **grad** u may be defined as

$$\mathbf{grad}\ u = \mathbf{i}\frac{\partial u}{\partial x} + \mathbf{j}\frac{\partial u}{\partial y} + \mathbf{k}\frac{\partial u}{\partial z} \qquad (D.8)$$

The scalar product of these two vectors is seen to be equal to du

$$du = (\mathbf{grad}\ u) \cdot \mathbf{dl} \qquad (D.9)$$

The partial derivatives are assumed to exist and be uniquely determined at the point. Then for a given infinitesimal magnitude of displacement the change in u is greatest when the angle between **grad** u and **dl** is least, *i.e.*, when **dl** is in the direction of the vector **grad** u. Thus **grad** u is seen to be a vector in the direction of the greatest space rate of change of u which is the reason for the name gradient or **grad** of u. A vector field which may be written as the gradient of a scalar field has an interesting and important property. From Eq. (D.9) the line integral of (**grad** u) · **dl** along the curve S as in Fig. (D.8) is simply the difference between the values of the scalar u at the two end points, *i.e.*,

$$\int_A^B (\mathbf{grad}\ u) \cdot \mathbf{dl} = \int_A^B du = u_B - u_A$$

This is evidently independent of the path followed. If the circuit is completed by returning to the point A by any other path, say S', the contribution along this path is the negative of the above quantity. Hence

$$\oint (\mathbf{grad}\ u) \cdot \mathbf{dl} = 0 \qquad (D.10)$$

Thus, if a vector field can be represented as the gradient of a scalar-point function, the value of its line integral over a closed path vanishes. One of the most important uses of the gradient is in connection with a single-valued potential-energy field such as an electrostatic or gravitational potential. If $\mathbf{F}$ is the force on a body in such a field and dW is the work done on the body in an infinitesimal displacement, **dl** resulting in a decrease in energy $-dU$ of the field, the following equality results:

$$dW = \mathbf{F} \cdot \mathbf{dl} = -dU = -(\mathbf{grad}\ U) \cdot \mathbf{dl}$$

Therefore

$$\mathbf{F} = -\mathbf{grad}\ U \qquad (D.11)$$

The force vector is determined immediately from the scalar-potential field U. For a finite displacement the work done is equal to the difference between the initial and final potential energies

$$W = \int_A^B \mathbf{F} \cdot \mathbf{dl} = -\int_A^B (\mathbf{grad}\ U) \cdot \mathbf{dl} = -\int_A^B dU = U_A - U_B$$

Divergence of a Vector Field.—It is evident that Eq. (D.8) can be formally regarded as the product of a vector $\mathbf{i}\,\partial/\partial x + \mathbf{j}\,\partial/\partial y + \mathbf{k}\,\partial/\partial z$ with a scalar u. The order of these factors is obviously important for the product $(\partial/\partial x)u$ means the partial derivative of u with respect to x. The above vector is known as the *nabla* operator and is written ∇. Thus **grad** u may be written ∇u (read nabla u). If $\mathbf{v}$ represents a vector field whose components together with their partial derivatives are continuous, a significant and very useful scalar field may be derived from it

by scalar multiplication of ∇ and $\mathbf{v}$. The scalar quantity resulting is known as the divergence of $\mathbf{v}$ or div $\mathbf{v}$.

$$\text{div } \mathbf{v} = \nabla \cdot \mathbf{v} = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right)(\mathbf{i}v_x + \mathbf{j}v_y + \mathbf{k}v_z)$$

Performing the indicated multiplication with the conventions of Eq. (*D*.2)

$$\text{div } \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \tag{D.12}$$

If the divergence of a vector is zero, the vector is said to be *solenoidal.* The significance of the concept of divergence may be seen from a consideration of the infinitesimal volume element in Fig. *D*.9. The value of the vector, representing, say, the flow of a fluid or an electric current, at the center of the element is $\mathbf{v}$. The central arrow represents its x component of magnitude v_x. By Taylor's theorem the values of the x component at the centers of the left- and right-hand faces are $v_x - \frac{\partial v_x}{\partial x}\frac{dx}{2}$, and $v_x + \frac{\partial v_x}{\partial x}\frac{dx}{2}$, respectively, to the first order of small quantities. Multiplying by the face areas, the net outward flow through this pair of faces is seen to be

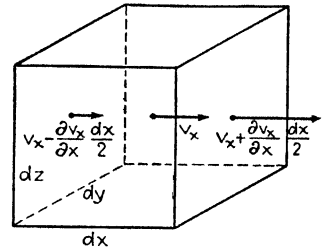



FIG. *D*.9.—Derivation of the theorem of flux.

$$-\left(v_x - \frac{\partial v_x}{\partial x}\frac{dx}{2}\right)dy\,dz + \left(v_x + \frac{\partial v_x}{\partial x}\frac{dx}{2}\right)dy\,dz = \frac{\partial v_x}{\partial x}dx\,dy\,dz$$

The other two pairs of faces may be treated in an analogous manner and after adding these three results, the net outward flow from the infinitesimal volume becomes

$$\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right)dx\,dy\,dz = \text{div } \mathbf{v}\,dv$$

where dv is written for the volume element $dx\,dy\,dz$. Thus the divergence of $\mathbf{v}$ times dv represents the net outward flow through the surface bounding the volume dv. A finite volume V may be considered to be made up of a large number of these infinitesimal volumes closely packed together. The algebraic sums of the flows across the internal surfaces of contact of these infinitesimal volume elements evidently vanish. Therefore the integral of div $\mathbf{v}\,dv$ throughout the volume represents the flow through

the external surfaces of these infinitesimal elements and the sum of these surfaces constitutes the bounding surface S of the volume V. Hence

$$\int_v \text{div } \mathbf{v} \, dv = \oint_s \mathbf{v} \cdot d\mathbf{s} \qquad (D.13)$$

This general result is known as the *theorem of flux* or *Gauss's theorem* It is of great value in the discussion of electric and magnetic fields.

Rotation or **curl** *of a Vector Field.*—If **v** represents a continuous differentiable vector field, a second vector field of great importance in electromagnetic theory may be derived from it by forming the vector product of ∇ and **v**. This product is known as **curl v**. It may be expressed in terms of its components by writing out the vectors ∇ and **v** and performing the indicated multiplication with the vector-product conventions given by Eq. (D.5).

$$\mathbf{curl}\ \mathbf{v} = \nabla \times \mathbf{v} = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right) \times (\mathbf{i}v_x + \mathbf{j}v_y + \mathbf{k}v_z)$$

$$\mathbf{curl}\ \mathbf{v} = \mathbf{i}\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right) + \mathbf{j}\left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}\right) + \mathbf{k}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right) \qquad (D.14)$$

This is also seen to be the expansion of the following determinant:

$$\mathbf{curl}\ \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_x & v_y & v_z \end{vmatrix} \qquad (D.14')$$

The vector **curl v** is always solenoidal, *i.e.*, div **curl v,** is equal to zero for any vector **v**. This may be seen in terms of the operator nabla and the triple-product equalities.

$$\nabla \cdot (\nabla \times \mathbf{v}) = (\nabla \times \nabla) \cdot \mathbf{v} = 0 \qquad \text{or} \qquad \text{div}\ \mathbf{curl}\ \mathbf{v} = 0$$

For the vector product of a vector with itself is zero from Eq. (D.6). If the vector **curl v** is zero, the vector **v** is said to be *irrotational*. A vector which may be represented as the gradient of a scalar field is always irrotational. Since **grad** u is ∇u, it follows immediately that

$$\mathbf{curl}\ (\mathbf{grad}\ u) = 0$$

The value of the concept of the **curl** of a vector may be seen by considering the line integral of a vector field around a closed curve. Consider an infinitesimal element of area in the xy plane. Let **F** be the value of a vector force field at the center of this element and evaluate the work done in proceeding around this element in the direction of the arrows of Fig. D.10. The component of the force parallel to y

at the right-band boundary is $F_y + \frac{\partial F_y}{\partial x}\frac{dx}{2}$, and that along the left-hand boundary is $F_y - \frac{\partial F_y}{\partial x}\frac{dx}{2}$. Therefore the contribution to the total work made on traversing the distance $+dy$ on the right side and $-dy$ on the left is $\partial F_y/\partial x \, dx \, dy$. In a similar way the contribution made by the top and bottom segments is $-\frac{\partial F_x}{\partial y} dx \, dy$. The total work done in circumscribing this elementary area is thus

$$\left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx \, dy = \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) ds_{xy}$$

Here ds_{xy} is written for an element of area in the xy plane. If a similar

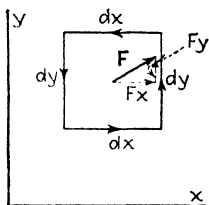



FIG. D.10.—Derivation of Stokes's theorem in two dimensions.

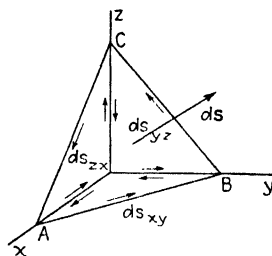



FIG. D.11.—Extension of the derivation of Stokes's theorem to three dimensions.

process is carried out in the yz and zx planes, the infinitesimal elements of work are found in an analogous manner to be

$$\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right) ds_{yz}$$

and

$$\left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right) ds_{zx}$$

respectively. Now consider an element of area ds whose normal is arbitrarily oriented. Let its projections on the three coordinate planes ds_{yz}, ds_{zx}, ds_{xy} be the triangular elements of area of Fig. D.11. Since these areas are equal to the area ds times the cosine of the angle between the normal to ds and the third axis, they are the components of the vector **ds** representing the area, *i.e.*,

$$\mathbf{ds} = \mathbf{i} \, ds_{yz} + \mathbf{j} \, ds_{zx} + \mathbf{k} \, ds_{xy}$$

On using Eq. D.14 and the above expression to form the scalar product of **curl F** and **ds**, the three terms which occur are seen to be the three elements of work done on traversing the three triangular elements of area in the senses indicated by the arrows of Fig. D.11. It is evident,

however, from this figure that the contiguous portions of these triangles lying along the axes are traversed twice, once in each direction. Therefore these contributions vanish and the total work is equivalent to that in circumscribing the element of area ds along the path $A \rightarrow B \rightarrow C$.

$$dW_{ds} = (\textbf{curl } \mathbf{F}) \cdot \mathbf{ds}$$

A finite area of a surface bounded by a closed curve can be subdivided into a large number of these infinitesimal areas. It may be seen from Fig. *D*.12 that the sum of the elements of work done in circumscribing all of these elements of area in the prescribed sense is the same as that done in traversing the bounding curve in the same sense. Thus the surface integral of **curl F** · **ds** over the surface is equal to the line integral of **F** · **dl** around the boundary.

FIG. *D*.12.—Extension of the circulation theorem to a finite surface.

$$\oint \mathbf{F} \cdot \mathbf{dl} = \int_s \textbf{curl } \mathbf{F} \cdot \mathbf{ds} \qquad (D.15)$$

This is known as *Stokes's theorem* and is of great importance in the theory of electromagnetism.

Useful Vector Relations Involving the Vector ∇.—The expression div **grad** u frequently occurs in electric and magnetic theory. It is evidently $\nabla \cdot \nabla u$ or $\nabla^2 u$. The scalar operator ∇^2 is known as *Laplace's operator* or the *Laplacian*.

$$\text{div } \textbf{grad } u = \nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \qquad (D.16)$$

The following vector identities may be established by expanding ∇ and the other vectors concerned in terms of their components:

$$\text{div } u\mathbf{A} = u \text{ div } \mathbf{A} + \mathbf{A} \cdot \textbf{grad } u \qquad (D.17)$$

$$\textbf{curl } u\mathbf{A} = u \textbf{ curl } \mathbf{A} - \mathbf{A} \times \textbf{grad } u \qquad (D.18)$$

$$\text{div } (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \textbf{curl } \mathbf{A} - \mathbf{A} \cdot \textbf{curl } \mathbf{B} \qquad (D.19)$$

$$\textbf{curl curl } \mathbf{A} = \textbf{grad} \text{ div } \mathbf{A} - \nabla^2 \mathbf{A} \qquad (D.20)$$

$$\textbf{grad } (\mathbf{A} \cdot \mathbf{B}) = \mathbf{A} \times \textbf{curl } \mathbf{B} + (\mathbf{A} \cdot \textbf{grad}) \mathbf{B} + \mathbf{B} \times \textbf{curl } \mathbf{A} + (\mathbf{B} \cdot \textbf{grad}) \mathbf{A} \qquad (D.21)$$

$$\textbf{curl } (\mathbf{A} \times \mathbf{B}) = (\mathbf{B} \cdot \textbf{grad})\mathbf{A} - (\mathbf{A} \cdot \textbf{grad})\mathbf{B} + \mathbf{A} \text{ div } \mathbf{B} - \mathbf{B} \text{ div } \mathbf{A} \qquad (D.22)$$

Here the vector $(\mathbf{A} \cdot \textbf{grad})\ \mathbf{B}$ stands for the vector

$$(\mathbf{A} \cdot \textbf{grad})\ \mathbf{B} = \mathbf{i}\left(A_x \frac{\partial B_x}{\partial x} + A_y \frac{\partial B_x}{\partial y} + A_z \frac{\partial B_x}{\partial z}\right) + \mathbf{j}\left(A_x \frac{\partial B_y}{\partial x} + A_y \frac{\partial B_y}{\partial y} + A_z \frac{\partial B_y}{\partial z}\right) + \mathbf{k}\left(A_x \frac{\partial B_z}{\partial x} + A_y \frac{\partial B_z}{\partial y} + A_z \frac{\partial B_z}{\partial z}\right)$$

in Cartesian coordinates.

Vector equations are quite independent of the coordinate system used. In the preceding discussion where vector components have occurred they have been expressed in terms of Cartesian coordinates.

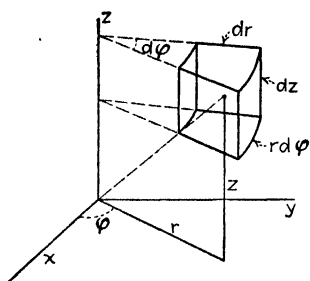



FIG. D.13.—Elements of length, area, and volume in cylindrical coordinates.

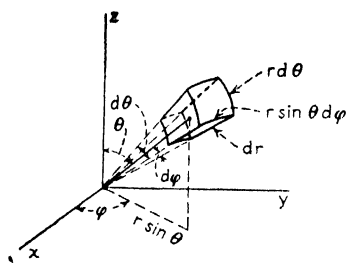



FIG. D.14.—Elements of length, area, and volume in spherical coordinates.

Many calculations are greatly facilitated by a suitable choice of coordinate system; the only special systems employed in this book are the cylindrical and spherical systems illustrated in Figs. D.13 and D.14. Vector quantities and components expressed in terms of either of these systems as listed below can be obtained from the fundamental vector concepts by using the line, surface, and volume elements appropriate to the particular system.

CYLINDRICAL COORDINATES

Orthogonal Line Elements

$dr,\ r\,d\phi,\ dz,$

Components of the Gradient

$$\text{grad}_r\, u = \frac{\partial u}{\partial r}$$

$$\text{grad}_\varphi\, u = \frac{1}{r}\frac{\partial u}{\partial \varphi}$$

$$\text{grad}_z\, u = \frac{\partial u}{\partial z}$$

Components of the Curl

$$\text{curl}_r\, \mathbf{A} = \frac{1}{r}\frac{\partial A_z}{\partial \varphi} - \frac{\partial A_\varphi}{\partial z}$$

$$\text{curl}_\varphi\, \mathbf{A} = \frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r}$$

$$\text{curl}_z\, \mathbf{A} = \frac{1}{r}\left[\frac{\partial}{\partial r}(rA_\varphi) - \frac{\partial A_r}{\partial \varphi}\right]$$

$$\text{div}\, \mathbf{A} = \frac{1}{r}\frac{\partial}{\partial r}(rA_r) + \frac{1}{r}\frac{\partial A_\varphi}{\partial \varphi} + \frac{\partial A_z}{\partial z}$$

Laplacian

$$\nabla^2 u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \varphi^2} + \frac{\partial^2 u}{\partial z^2}$$

SPHERICAL COORDINATES

Orthogonal Line Elements

$dr,\ r\,d\theta,\ r \sin\theta\, d\phi$

Components of the Gradient

$$\text{grad}_r\, u = \frac{\partial u}{\partial r}$$

$$\text{grad}_\theta\, u = \frac{1}{r}\frac{\partial u}{\partial \theta}$$

$$\text{grad}_\varphi\, u = \frac{1}{r\sin\theta}\frac{\partial u}{\partial \varphi}$$

Components of the Curl

$$\text{curl}_r\, \mathbf{A} = \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}(\sin\theta\, A_\varphi) - \frac{\partial A_\theta}{\partial \varphi}\right]$$

$$\text{curl}_\theta\, \mathbf{A} = \frac{1}{r\sin\theta}\frac{\partial A_r}{\partial \varphi} - \frac{1}{r}\frac{\partial (rA_\varphi)}{\partial r}$$

$$\text{curl}_\varphi\, \mathbf{A} = \frac{1}{r}\left[\frac{\partial (rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial \theta}\right]$$

$$\text{div}\, \mathbf{A} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2A_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(\sin\theta\, A_\theta) + \frac{1}{r\sin\theta}\frac{\partial A_\varphi}{\partial \varphi}$$

Laplacian

$$\nabla^2 u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 u}{\partial \varphi^2}$$

E. Units and Standards.—The formulation of experimental results in a new field is qualitative until a system of units is introduced. For instance, Coulomb found that the force between two charges was proportional to the magnitude of each of the charges and inversely proportional to the square of their separation. Likewise Ampère observed that force between two circuits carrying currents was proportional to the magnitudes of the currents and a function of the linear dimensions of the circuits and their separation. These experimental observations not only form the basis of the subject but also yield the relations by means of which a system of units can be defined. Equations may be set up by introducing a constant of proportionality into these relations. Since this process is entirely arbitrary, the constant is generally chosen in such a way as to simplify the type of calculation most frequently performed. For instance, in the field of electrostatics a typical calculation is that of the force between point charges in free space. Hence, for simplicity the constant in the Coulomb law is chosen as unity and the equation becomes

$$F = \frac{q_1 q_2}{r^2}$$

This is the basis of the electrostatic system of units (esu.). The mechanical units are the centimeter, gram, and second; a unit charge is of such a magnitude that when placed 1 cm. from an exactly equal charge in free space, it is repelled by it with a force of one dyne. Similarly, for calculations in the field of magnetostatics it is most convenient to define a unit pole in an analogous way. However, since electric and magnetic phenomena are interrelated, as the theory is developed, equations, relating charges and currents or poles are obtained and these are not in a particularly convenient form for calculation. Hence, for electromagnetic work the units are more conveniently defined in a different way. Incidentally, in the field of atomic physics these units are all too large for convenience. That type of calculation is facilitated by choosing as units of mass, charge, length, and time the mass and charge of the electron and the radius and period of rotation associated with the first Bohr orbit in hydrogen.

It is not possible to choose a system of units that is ideally suited to all types of calculation, but it is desirable for the purpose of consistency to adopt one system of units for use throughout. The internationally accepted absolute practical system of units has here been chosen and all quantities appear in these units unless otherwise specified. It is the most generally accepted system in use at present and has many convenient features to recommend it. In general electrical work the Coulomb laws are not used as frequently as are the electromagnetic relations.

Hence the system is chosen in such a way as to simplify these latter expressions, namely,

$$\text{curl } \mathbf{H} = \mathbf{i}_v \qquad \text{and} \qquad \text{curl } \mathbf{E} = -\frac{d\mathbf{B}}{dt}$$

Also the units of ampere, ohm, volt, etc., which are of convenient magnitudes and have been widely used for many years are retained. Likewise the mechanical units of work and power are the familiar joule and watt. The two actual units of mass and length (the kilogramme des Archives and the international prototype meter) are chosen as the fundamental standard units of these quantities. The unit of time is the second (1/86,400 of a mean solar day). No constants of proportionality appear in the ordinary mechanical equations. The unit of force, which is called the newton, is that force which applied to a mass of 1 kilogram will impart to it an acceleration of 1 meter per second per second. It is of a more convenient order of magnitude for most work than the dyne. The joule and joule per second or watt develop in the logical way. Two constants of proportionality, however, do appear in the electrostatic and magnetostatic Coulomb laws; just as a constant appears in the law of gravitation since all the quantities occurring have been previously defined. These are no more inconvenient than the 4π and c (velocity of light) that appear in the Gaussian system, for instance. Also, they serve in a sense to distinguish between the two types of electric and magnetic vectors, namely, **E** and **D** and **B** and **H.**

The technique of measuring electrical quantities in terms of the fundamental standards of length, mass, and time has developed to such a point that it is now feasible to define the electrical quantities in terms of these standards. The quantities defined in this way are known as *absolute practical* units. By international agreement a dimensional quantity $4\pi \times 10^{-7}$ henrys per meter (μ_0) is assigned to the permeability of free space. By an experiment such as that of Lorenz (Sec. 10.4) the emf. $\mathcal{E}_1$ developed in a circuit by changing its mutual inductance with respect to a second circuit carrying a current i_2 can be accurately compared with the potential drop in a resistance R carrying the current i_2 or some known fraction of it. In this case $\mathcal{E}_1 = i_2\, \partial L_{12}/\partial t$ and $\mathcal{E}_1' = i_2R$, hence R can be determined in terms of the second and L_{12} which involves only the meter and μ_0. The unit of current, the ampere, can be defined either in terms of the force between circuits carrying currents or in terms of the joule heating in a known resistance. The volt is then defined as the emf. or potential difference between the terminals of a 1-ohm resistance carrying a current of 1 amp. The amount of electricity transported per second under these circumstances is 1 coulomb. This definition of the coulomb determines the constant appearing in the electrostatic law

of force, *i.e.*, κ_0 the permittivity of free space. Hence a unit electric field is one in which a charge of 1 coulomb experiences a force of 1 newton. The coefficient of self- or mutual inductance can in simple cases be calculated from μ_0 and the geometry, otherwise it is measured by the force between two circuits carrying known currents or by the emf. induced by a known rate of change of current. If a loop of wire has a self-inductance of 1 henry, it is threaded by 1 weber of flux when carrying a current of 1 amp. The unit of magnetic induction is the weber per square meter. If the force per meter of length acting on a wire carrying a current of 1 amp. perpendicular to the lines of magnetic induction is 1 newton, the value of the magnetic induction is 1 weber per square meter. Also a loop of wire 1 $m.^2$ in area carrying a current of 1 amp. has a unit magnetic moment. An ideal permanent magnet 1 m. long having the same moment has unit poles. This definition is in agreement with the Coulomb law of magnetostatics if the constant of proportionality there appearing is $1/(4\pi\mu_0)$. The concept of reluctance is convenient when dealing with electromagnets. It is the analogue of resistance in an ohmic circuit. An iron core which has a unit reluctance of 1 amp.-turn produces a flux of 1 weber. The constants κ_0 and μ_0 are related experimentally through the velocity of propagation of electromagnetic radiation, $c = 1/(\kappa_0\mu_0)^{1/2}$.

The legal units of the ohm and the ampere in use prior to 1948 differed slightly from this absolute system. These international or legal units were defined as follows:

1 international ohm is the resistance offered to an unvarying electric current by a column of mercury at the temperature of melting ice 14.4521 gm. in mass, of constant cross-sectional area and 106.3 cm. long.

1 international ampere is the value of the unvarying current which on passing through a solution of silver nitrate in water in accordance with standard specifications deposits silver at the rate of 0.001118 gm. per second.

These definitions were never very satisfactory and now that measurements in terms of the kilogram, meter, and second can be made with precision, they have been replaced by the units of the absolute system outlined in the preceding paragraph. The most recent experimental relations between these two systems are given below:

1 international ohm = 1.00049 absolute ohms
1 international ampere = 0.999835 absolute ampere

From these it is evident that 1 international volt is equal to 1.00034 absolute volts, 1 international joule is equal to 1.000165 absolute joules, etc.

Precision standardization in terms of the fundamental units of length, mass, and time is carried out only in primary standardizing laboratories

such as the Bureau of Standards or the National Physical Laboratory.[1] In these institutions standard ohms are prepared and calibrated by the Lorenz type of technique. Currents, measured with a current balance (Sec. 9.5), are used in conjunction with standard ohms to calibrate standard cells. The standard ohm and standard cell are then used as the basic standards of resistance and emf. in ordinary laboratories. The standard cell is used with a potentiometer and volt box to calibrate voltmeters. Ammeters are calibrated by means of a set of standard resistances together with a standard cell and potentiometer (Sec. 4.7). In addition, a laboratory generally possesses standards of capacity and self- and mutual inductance. With this equipment all the ordinary types of electrical measurements can be readily performed.

Certain of the more fundamental equations relating electrical quantities in the system of units here adopted are collected below for reference.

ELECTROSTATICS

$$\mathbf{F} = \frac{1}{4\pi\kappa\kappa_0}\frac{q_1q_2}{r_{12}^3}\mathbf{r} = q\mathbf{E} \qquad V = \int \mathbf{E}\cdot d\mathbf{l} = \frac{1}{4\pi\kappa\kappa_0}\int\frac{q_v}{r}dv \qquad \mathbf{E} = -\mathbf{grad}\ V$$

$$q = \int q_v\,dv \qquad q = VC \qquad \mathbf{curl}\ \mathbf{E} = 0 \qquad \mathrm{div}\ \mathbf{E} = \frac{q_v^t}{\kappa_0} \qquad \nabla^2 V = -\frac{q_v^t}{\kappa_0}$$

$$U' = -\mathbf{p}\cdot\mathbf{E} = qV \qquad \mathbf{F} = (\mathbf{p}\cdot\mathbf{grad})\mathbf{E} \qquad \mathbf{T} = \mathbf{p}\times\mathbf{E}$$

$$\mathbf{p} = \int \mathbf{p}_v\,dv \qquad \mathbf{p}_v = \chi_e\kappa_0\mathbf{E} \qquad \mathrm{div}\ \mathbf{p}_v = -q_v^i \qquad \mathbf{p}\cdot\mathbf{n} = q_s^i$$

$$\mathbf{D} = \kappa_0\mathbf{E} + \mathbf{p}_v = \kappa\kappa_0\mathbf{E} \qquad \mathrm{div}\ \mathbf{D} = q_v$$

$$U = \tfrac{1}{2}\sum_i q_iV_i = \tfrac{1}{2}\sum_i\sum_j C_{ij}V_iV_j = \tfrac{1}{2}\int q_vV\,dv = \tfrac{1}{2}\int \mathbf{D}\cdot\mathbf{E}\,dv$$

STATIC CURRENTS

$$\mathrm{div}\ \mathbf{i}_v = 0 \qquad i = \int \mathbf{i}_v\cdot d\mathbf{s} \qquad \mathbf{i}_v = \sigma\mathbf{E} \qquad V = \mathcal{E} = iR$$

$$R = \frac{\kappa_0}{C_0\sigma} \qquad P = \mathcal{E}i = \int \mathbf{E}\cdot\mathbf{i}_v\,dv$$

STATIC ELECTROMAGNETISM

$$d\mathbf{F}_1 = \frac{\mu\mu_0}{4\pi}i_1i_2\frac{d\mathbf{l}_1\times(d\mathbf{l}_2\times\mathbf{r})}{r_{12}^3} = i_1d\mathbf{l}_1\times\mathbf{B}$$

$$\phi = \int \mathbf{B}\cdot d\mathbf{s} \qquad \mathbf{B} = \mathbf{curl}\ \mathbf{A} \qquad \mathrm{div}\ \mathbf{B} = 0 \qquad \mathbf{curl}\ \mathbf{B} = \mu_0\mathbf{i}_v^t$$

$$\nabla^2\mathbf{A} = -\mu_0\mathbf{i}_v^t \qquad \mathrm{div}\ \mathbf{A} = 0 \qquad \mathbf{A} = \frac{\mu\mu_0}{4\pi}\int\frac{\mathbf{i}_v\,dv}{r} = \frac{\mu\mu_0}{4\pi}i\oint\frac{d\mathbf{l}}{r}$$

$$U' = \mathbf{m}\cdot\mathbf{H} = i\varphi \qquad \mathbf{F} = (\mathbf{m}\cdot\mathbf{grad})\,\mathbf{H} \qquad \mathbf{T} = \mathbf{m}\times\mathbf{H}$$

$$\mathbf{m} = \int \mathbf{m}_v\,dv = \mu_0 i\int d\mathbf{s} \quad \mathbf{m}_v = \chi_m\mu_0\mathbf{H} \quad \mathbf{curl}\ \mathbf{m}_v = \mu_0\mathbf{i}_v^a \quad \mathbf{m}_v\times\mathbf{n} = \mu_0\mathbf{i}_s^a$$

[1] BRIGGS, *Rev. Mod. Phys.*, **11**, 111 (1939).

$$\text{curl } \mathbf{H} = \mathbf{i}_v \qquad \mu_0\mathbf{H} = \mathbf{B} - \mathbf{m}_v = \frac{\mathbf{B}}{\mu} \qquad L_{12} = \frac{\mu\mu_0}{4\pi}\oint\oint\frac{d\mathbf{l}_1 \cdot d\mathbf{l}_2}{r_{12}}$$

$$U = \tfrac{1}{2}\sum_i i_i\phi_i = \tfrac{1}{2}\sum_i\sum_j L_{ij}i_i i_j = \tfrac{1}{2}\int \mathbf{i}_v \cdot \mathbf{A}\, dv = \tfrac{1}{2}\int \mathbf{B} \cdot \mathbf{H}\, dv$$

VECTORS FUNCTIONS OF THE TIME

$$\text{curl } \mathbf{H} = \mathbf{i}_v + \frac{\partial \mathbf{D}}{\partial t} \qquad \text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \mathcal{E} = -\frac{\partial \phi}{\partial t}$$

$$\text{div } \mathbf{i}_v = -\frac{\partial q_v}{\partial t} \qquad -\mathbf{E} = \frac{\partial \mathbf{A}}{\partial t} + \textbf{grad } V$$

$$\mathbf{F} = q(\mathbf{E} + \mathbf{u} \times \mathbf{B}) \qquad \mathbf{N} = \mathbf{E} \times \mathbf{H} \qquad c = \frac{1}{(\kappa_0\mu_0)^{1/2}}$$

$$-\nabla^2\mathbf{A} + \frac{\kappa\mu}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = \mu\mu_0\mathbf{i}_v \qquad -\nabla^2 V + \frac{\kappa\mu}{c^2}\frac{\partial^2 V}{\partial t^2} = \frac{q_v}{\kappa\kappa_0}$$

Analogous equations for **B**, **H**, **E**, and **D**.

DIPOLE RADIATION

$$d\mathbf{B} = \frac{\mu_0}{2r^2\lambda} i e^{2\pi j\left(vt - \frac{r}{\lambda}\right)} d\mathbf{l} \times \mathbf{r} \qquad d\mathbf{E} = \frac{1}{(\kappa_0\mu_0)^{1/2}} d\mathbf{B} \times \mathbf{r}_1$$

$$d\mathbf{E} = \frac{(\mu_0/\kappa_0)^{1/2}}{2r^2\lambda} i e^{2\pi j\left(vt - \frac{r}{\lambda}\right)} (d\mathbf{l} \times \mathbf{r}) \times \mathbf{r}_1$$

For an explanation of vector nomenclature see Appendix *D*. The dielectric constant κ, and the permeability μ, are pure numbers. The subscripts v and s indicate that the quantity is taken per unit volume or per unit surface area. The subscript 1 means a unit vector, or 1 and 2 distinguish between two quantities. The superscripts indicate the categories into which charges and currents are divided for convenience: q^t total charge, q^i induced charge (in polarization), q free charge, i^t total current, i^a amperian currents (in magnetization), i free current.

NUMERICAL VALUES OF ATOMIC AND ELECTRIC CONSTANTS[1]

1. Velocity of light $c = (\kappa_0\mu_0)^{-1/2} = (2.99776 \pm 0.00004) \times 10^8$ m./sec.
2. Impedance of free space $(\mu_0/\kappa_0)^{1/2} = 376.707 \pm 0.004$ ohm
3. Permittivity of free space $\kappa_0 = (8.85525 \pm 0.00025) \times 10^{-12}$ farad/m.
4. Permeability of free space $\mu_0 = 4\pi \times 10^{-7}$ henry/m.
   $= 1.257 \times 10^{-6}$ henry/m.
5. Faraday constant $= F = (96514.0 \pm 10)$ coulombs/gm. equivalent ($0^{16} = 16$ physical scale)
6. Avogadro number $= N = (6.0228 \pm 0.0011) \times 10^{23}$ 1/mole
7. Boltzmann constant $= k = (1.38047 \pm 0.00026) \times 10^{-23}$ joule/°C.
8. Planck constant $= h = (6.6242 \pm 0.0024) \times 10^{-34}$ joule sec.
9. Electronic charge $= e = (1.60203 \pm 0.00034) \times 10^{-19}$ coulomb
10. Specific electronic charge $= e/m = (1.7592 \pm 0.0005) \times 10^{11}$ coulombs/kg.
11. Electronic magnetic moment $= m_B = (0.9273 \pm 0.0003)$ $\times 10^{-23}$ joule m./amp. turn

[1] BIRGE, *Rev. Mod. Phys.*, **13**, 233 (1941).

RELATIONS BETWEEN THE PRINCIPAL ELECTRIC AND MAGNETIC QUANTITIES IN THE DIFFERENT SYSTEMS OF UNITS

Equality signs are implied across any row, *i.e.*, 1 m. = 100 cm.

Entity	Symbol	Absolute practical	Electrostatic	Electromagnetic
Length	l	1 meter	100 centimeters	100 centimeters
Mass	m	1 kilogram	1,000 grams	1,000 grams
Time	t	1 second	1 sec.	1 sec.
Force	$\mathbf{F}$	1 newton	10^5 dynes	10^5 dynes
Work Energy	W U	1 joule	10^7 ergs	10^7 ergs
Power	P	1 watt	10^7 ergs/sec.	10^7 ergs/sec.
Charge	q	1 coulomb	3×10^9	10^{-1}
Current	i	1 ampere	3×10^9	10^{-1}
Electric field	$\mathbf{E}$	1 volt/meter	$1/(3 \times 10^4)$	10^6
Electromotive force or potential difference	$\mathcal{E}$ V	1 volt	$\frac{1}{300}$	10^8
Polarization	$\mathbf{p}_v$	1 coulomb/meter²	3×10^5	10^{-5}
Displacement	$\mathbf{D}$	1 coulomb/meter²	$12\pi \times 10^5$	$4\pi \times 10^{-5}$
Conductivity	σ	1 mho/meter	9×10^9	10^{-11}
Resistance	R	1 ohm	$1/(9 \times 10^{11})$	10^9
Capacity	C	1 farad	9×10^{11}	10^{-9}
Flux	φ	1 weber	$\frac{1}{300}$	10^8 maxwells
Magnetic induction	$\mathbf{B}$	1 weber/meter²	$1/(3 \times 10^6)$	10^4 gausses
Magnetic field	$\mathbf{H}$	1 ampere turn/meter	$12\pi \times 10^7$	$4\pi \times 10^{-8}$ oersted
Magnetomotive force	$\mathcal{H}$	1 ampere turn	$12\pi \times 10^9$	$4\pi/10$ gilbert
Magnetization	$\mathbf{m}_v$	1 weber/meter²	$1/(12\pi \times 10^6)$	$10^4/4\pi$
Inductance	L	1 henry	$1/(9 \times 10^{11})$	10^9
Reluctance	$\mathcal{R}$	1 ampere turn/weber	$36\pi \times 10^{11}$	$4\pi \times 10^{-9}$
Pole strength	p_m	1 weber	$1/(12\pi \times 10^2)$	$10^8/4\pi$
Permittivity of free space	κ_0	$1/(36\pi \times 10^9)$ farad/meter	unity	$1/(9 \times 10^{20})$ (sec./cm.)²
Permeability of free space	μ_0	$4\pi \times 10^{-7}$ henry/meter	$1/(9 \times 10^{20})$ (sec./cm,)²	unity

The velocity of electromagnetic radiation is taken for simplicity as

$$c = 3 \times 10^8 \text{ meters/second.}$$

INDEX

A

N